

XEROX
COPY

DO NOT
GIVE AWAY

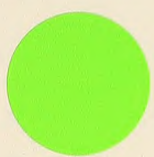


TABLE OF CONTENTS

Issues Related to the Volume and Intensity
of Physician Services

by Mark V. Pauly, Project Director
John Eisenberg, Principal Investigator
Roger Feldman, Principal Investigator
J. Sanford Schwartz, Alan L. Hillman, M. Haim Erder
James Highland, Margaret Higgins Radany, Eugene Rich

Federal Project Officer: Sherry Terrell

Leonard Davis Institute of Health Economics
University of Pennsylvania

Division of Health Services Research and Policy
University of Minnesota

HCFA Contract No. 99-C-99169/5-01

December 21, 1988

NTIS NO. PB 89 151591

FINAL
DEC 21 1988

TABLE OF CONTENTS

	<u>Page</u>
1.0. EXECUTIVE SUMMARY.....	1
1.1. Purpose and Scope of the Study.....	1
1.2. Models of Physician Behavior.....	2
1.3. Likely Effects of Volume Controls.....	3
1.4. Use of Volume Controls by Medicare and Commercial Carriers.....	7
1.5. Volume Effects of the RBRVS.....	9
2.0. INTRODUCTION.....	10
2.1. Factors Effecting Volume and Intensity.....	11
3.0. MODELS OF PHYSICIAN BEHAVIOR.....	14
3.1. Profit Maximization.....	14
3.1.1. Description of Model.....	14
3.1.2. Profit Maximization and the Effect of Price on Volume..	15
3.1.3. Adding Inducement.....	15
3.2. The Sophisticated Target Income Model.....	16
3.2.1. The Sophisticated Model.....	16
3.2.2. Determinants of Inducement.....	17
3.3. The Patient Agency Model.....	18
3.3.1. The Clinical Agent.....	19
3.3.2. The Economic Agent.....	19
3.3.3. Patient Preferences and Demand.....	20
3.3.4. Patient Convenience.....	20
3.3.5. Other Factors.....	20
4.0 DESCRIPTION OF CONTROLS.....	21
4.1. Clinical Guidelines With or Without Education.....	21
4.2. Utilization Review With or Without Penalties.....	21
4.3. Copayment.....	22
4.4. Capitation With or Without Financial Incentives.....	22
4.5. Expenditure Target.....	23
4.6. Collapsed Procedure Packages (Coding Changes).....	24
4.7. Service Packaging (Bundling Services).....	24
5.0 ANALYSIS OF CONTROLS.....	26
5.1. Clinical Guidelines With Professional Education.....	27
5.1.1. Overview of the Literature.....	27
5.1.2. Likely Effects of Implementation.....	28
5.1.3. Summary.....	37
Appendix: Theoretical Effects of Guidelines Under Behavioral Models.....	38

TABLE OF CONTENTS

(continued)

	<u>Page</u>
5.2. Utilization Review.....	43
5.2.1. Overview of the Literature.....	43
5.2.2. Likely Effects of Implementation.....	44
5.2.3. Summary and Conclusions for Retrospective Utilization Review.....	50
Appendix: Theoretical Effects of Utilization Under Behavioral Models.....	51
5.3. Copayments.....	57
5.3.1. Overview of the Literature.....	57
5.3.2. Likely Effects of Implementation.....	58
5.3.3. Summary.....	65
Appendix: Theoretical Effects of Copayment Under Behavioral Models.....	66
5.4. Capitation.....	71
5.4.1. Overview of the Literature.....	71
5.4.2. Likely Effects of Implementation.....	72
5.4.3. Summary.....	78
Appendix: Theoretical Effects of Capitation Under Behavioral Models.....	80
5.5. Expenditure Growth Targets.....	84
5.5.1. Overview of the Literature.....	84
5.5.2. Likely Effects of Implementation.....	84
5.5.3. Summary.....	99
Appendix: Theoretical Effects of Expenditures Under Behavioral Models.....	101
5.6. Collapsed Coding.....	107
5.6.1. Overview of the Literature.....	107
5.6.2. Likely Effects of Implementation.....	107
5.6.3. Summary.....	114
Appendix: Theoretical Effects of Collapsed Procedure Codes Under Profit Maximization.....	115
Appendix: The Relationship of the Number of Procedure Codes to Total Medicare Cost: A Numerical Example.....	119
5.7. Service Bundling.....	123
5.7.1. Overview of the Literature.....	123
5.7.2. Likely Effects of Implementation.....	123
5.7.3. Summary.....	129
Appendix: Theoretical Effects of Bundling Under Behavioral Models.....	130

TABLE OF CONTENTS

(continued)

	<u>Page</u>
6.0 SURVEY OF MEDICARE AND PRIVATE SECTOR CARRIERS.....	135
6.1. Methods.....	135
6.2. Results.....	136
6.2.1. Medicare Carriers.....	136
6.2.2. Survey of Commercial Health Insurance Carriers.....	143
6.3. Discussion.....	158
7.0. SUMMARY AND RECOMMENDATIONS.....	160
7.1. Three Models of Physician Behavior.....	160
7.2. Principles of Volume Control.....	161
7.3. Volume Controls.....	166
7.3.1. Guidelines.....	166
7.3.2. Utilization Review.....	168
7.3.3. Co-payment and Deductibles.....	169
7.3.4. Capitation.....	170
7.3.5. Expenditure Growth Targets.....	171
7.3.6. Collapsed Procedure Codes.....	173
7.3.7. Bundling of Procedures.....	174
7.4. Surveys of Commercial Insurance Firms and Medicare Carriers....	175
7.5. Impact of Resource-Based Relative Value Scale.....	176
REFERENCES.....	177
APPENDIX I: FINANCIAL INCENTIVES IN MEDICAL DECISION MAKING	
APPENDIX II: SURVEY INSTRUMENTS	

1.0. EXECUTIVE SUMMARY

1.1. Purpose and Scope of the Study

This study addresses a variety of issues related to the volume and intensity of physician services. Three major areas are explored. First, major emphasis has been placed on an evaluation of seven methods that might be used to control volume and intensity, including a determination and analysis of the "most likely" effects of each method were it implemented. These determinations are based on an analysis of economic theory, empirical evidence, and behavioral theories of physician behavior. Design issues that would affect the performance of the controls are also discussed.

The seven controls examined here include:

- o Clinical Guidelines with Professional Education
- o Utilization Review
- o Copayment
- o Capitation
- o Expenditure Cap
- o Collapsed Coding
- o Service Bundling

Because assumptions about the ways in which physicians make patient care decisions have a significant impact on predictions about the effects of these volume controls, the three prevailing theories of physician behavior; namely, profit maximization, target income and patient agency, are discussed in some detail early in the report.

The second concern of this study is the use of volume controls by Medicare and commercial insurance carriers. Through surveys done in collaboration with the Health Care Financing Administration, Bureau of Program Operations and the Health Insurance Association of American (HIAA), all Medicare carriers and 120 commercial carriers were surveyed regarding their methods of paying physicians and their use of specific volume controls. Perceptions of the effectiveness of these controls were also assessed.

Finally, a consideration of the impact of the Resource-Based Relative Value Scale (RBRVS) on the volume of physician services is presented. Because this portion of the study differs in style and substance, and because it is intended for a somewhat different audience, its results are reported in a separate technical report. Also, based on economic and behavioral theory, this part of the study presents the possible effects of the RBRVS under various assumptions of physician behavior and design features (e.g., with and without balance billing allowed). Analysis of these combinations of possibilities are intended to allow for knowledgeable decision making by policymakers.



1.2. Models of Physician Behavior

The first model of physician behavior is that of the Clinician as Patient's Agent. In the patient agency model, one assumes that the physician acts as the patient's advocate, and that this is the physician's primary or even only motivation. In this scenario, decisions are made by the physician in the patient's best interest; that is, the physician makes each decision based on what he or she believes the patient would want if the patient had the same information available to the physician. Our formulation of this model includes several components. First, the physician serves as the patient's perfect agent with regard to clinical outcome, thus attempting to optimize the patient's health for a given level of patient out-of-pocket expenditure. The second component of the Patient Agency Model is the physician serving as the patient's economic agent. Physicians become their patients' advocates with regard to the cost of medical care, attempting to avoid undue or avoidable out-of-pocket expenditures for the patient. Each of these components of the patient agency model requires that the physician understand the patient's utilities or values for various clinical outcomes and patients' disutilities for side effects, complications, and inconvenience or discomfort in undergoing a medical service. In addition to being influenced by the patient, the physician as the patient's agent will also be influenced by professional standards and preferred styles of practice. In the best case, these professional standards will hold each physician to a measure of quality expected by the profession, and to standards of professionalism and maintenance of expertise that society would expect of a self-policing and self-controlling profession. Other professional influences may be less idealized and more a function of the sociologic temper of the profession in a given area or specialty.

The role of the physician as agent of the patient may conflict with the second model of physician decision making, that of the physician as a Profit Maximizing Businessman. In this model, the physician will attempt to provide a set of services at a quantity which maximizes his or her net profit, given a current price and the opportunity cost of physician time, or to establish a price which will maximize income. Short-term desires of the profit maximizing physician to increase income will be tempered by long-term considerations, which require that the physician maintain an acceptable reputation and patient satisfaction in order to be able to command enough demand for his or her services in the future.

The third model of medical decision making explored in this analysis is that of the Sophisticated Physician's Target Income. In this model of physician behavior, physicians' utility from medical practice is a function of both the income that they can earn and the psychic cost of having to create demand for their own services. Therefore, a physician might seek a certain income, called the target income, but will do so only if the psychological expense of compromising personal or professional values would not be so large as to discourage the creation of demand for medical care. Therefore, in this model, induced demand is constrained by the subjective value that the physician attaches to practicing ideal medicine, either because of potential harm to the patient's utility, or because of concern about deviating from correct practice. As in the Profit Maximizing Model, induced demand will also

be constrained by long-term considerations regarding reputation and detectable indicators of quality.

Since none of these three models of physician behavior has been established empirically in the health economics or medical decision making literature, we believe that it is important to consider all three in evaluating the likely impact of a volume control measure. It is likely that different physicians will be influenced by the incentives reflected in these three models to different degrees, and that, as a whole, the profession is probably influenced to some degree by all of them.

1.3. Likely Effects of Volume Controls

Recommendations: Guidelines

Although guidelines in and of themselves would not be a strong volume control, we think they are important adjuncts to utilization review, expenditure caps, capitation and copayment. Thus, should any of these controls be implemented, we would strongly recommend the development and concurrent use of guidelines. The following points should be considered when developing or using guidelines:

- o Recent experience suggests that guidelines should be professionally derived, clearly defined, based upon data, and should demonstrate the possibility for reduction in cost while preserving or enhancing the quality of medical care or enhancing the quality of medical care at acceptable cost. Although it was felt in the 1970s that standards established by PSRO's should be local in nature, the wide variation in medical practices and the technical expertise necessary to develop guidelines has lead many experts to now believe that a set of standards should be established nationally with some room for local modification.
- o Guidelines are most likely to be effective if they are accompanied by incentives to physicians to adhere to the guidelines, or by sanctions for deviations.
- o The scientific basis for guidelines should be the same regardless of whether they are used for educational purposes (that is, as recommendations for idealized practice, or "pathways") or utilization review (that is, as criteria for negative feedback or penalties to the physicians, or "boundaries"). The design of guidelines, however, depends on their intended purpose. Specifically, guidelines developed for use in utilization review require stronger evidence before restricting reimbursement or imposing penalties on medical practices. The basis for guidelines (efficacy, safety, cost, available budget) should be specified, and can be different in different situations.
- o Guidelines should be widely disseminated, including to patients, who can then use them to evaluate the options available and to assess the

care being provided to them. Guidelines should accommodate informed differences in patient preferences for risk, amenity, or convenience of access.

- o Major public-sector initiatives are needed to develop improved methods of producing guidelines, to collect better data regarding the outcomes, risk and cost of medical care, and to facilitate or sponsor the development of sets of guidelines for services that are suspected of under- or overutilization. Funding for a national center or institute should be given serious consideration.
- o Evaluation of the effectiveness of guidelines in controlling volume and/or promoting appropriate use of medical services should be promoted and sponsored.
- o Data on outcomes, risk and cost should be required for new services when Medicare decides to reimburse for their provision, and guidelines suggesting appropriate use of the services should be made available.

Recommendations: Utilization Review

Cost-effective utilization review should be an element of any broader scheme of volume control, including the methods described in this report. In developing and implementing a UR system, the following should be considered:

- o Physicians should be involved in the process of utilization review and vested in its outcome, including participation in the design of UR criteria, involvement in its implementation, responsibility for appropriate appeals, and being influenced in a meaningful way by its findings (including the possibility of financial penalties).
- o Standards for UR should be unambiguous.
- o UR should be linked with assessment of the quality of care and designed to enhance quality.
- o UR should first be targeted to a few problems, especially those suspected to have large amounts of underuse or overuse.
- o Patients should be involved in the process and, ideally, should be affected financially by its results in order to gain their participation and interest. If they desire to override UR decisions, they should pay the cost.
- o UR should be accompanied by price adjustments to meet likely changes in demand for services resulting from the review process.
- o It is premature to initiate a full-blown UR program for Part B of Medicare. Criteria are needed; effective methods of influencing

physicians are needed; and prospective review procedures should be developed.

- o It is important to involve peers and leaders of the medical community.
- o UR programs should provide personal and personalized feedback to physicians.

Recommendations: Co-Payment/Deductibles

We believe that revitalizing the effect of the Part B copayment by discouraging the use of Medigap insurance is a good way to control volume and expenditures. This approach would be particularly appropriate when accompanied by a strong utilization review program based on clinical guidelines, to monitor for underutilization of services. Revitalization of copayments could be accomplished in the following ways:

- o Tax the value of employer-paid Medigap coverage as a part of retirees' income. This taxation might be targeted at Medigap policies to protect "appropriate coverage," for example for services thought to be underutilized, where underconsumption is feared, or for services that are not covered by Medicare.
- o Institute a surcharge on Medigap policies, for example added to the Part B premium and related to the beneficiaries' income or wealth.

Recommendations: Capitation

The administrative and design problems associated with capitation of Part B only make it a less attractive option (at this point in time) than expenditure caps with more rigorous utilization review or adjustments in copayment. To make capitation more feasible in Part B, the following should be undertaken:

- o Develop techniques to measure severity of disease or other predictors of medical care utilization in order to establish prices for capitated services.
- o Initiate and evaluate selected programs of Partial Physician Capitation, with groups of physicians chosen for capitation of physician services only on the basis of their site of care or the type of illness being treated.
- o Encourage case management techniques, which might be utilized in fee-for-service practice (especially in conjunction with utilization review) as well as in capitated practice.

Recommendations: Expenditure Targets

We believe that expenditure targets, especially in conjunction with rigorous utilization review based on sound clinical guidelines, offer an attractive approach to control of Part B expenditures. Expenditure targets are sure to control expenditures, and they provide Congress, HCFA and beneficiaries a predictability about Medicare expenses and premiums that is presently lacking. They also allow for a mechanism by which costs associated with changes in technology or in the health status of the Medicare population can be dealt with. We suggest that the following factors should be considered in the implementation of expenditure targets:

- o Appropriate populations of Medicare beneficiaries must be selected for inclusion. The following considerations enter into the decision about the geographic area covered by a target:
 1. The population covered should be large enough to be administratively feasible and to avoid a large number of small administrative units.
 2. The physician population within a target area should be small enough to allow for peer interaction and influence.
 3. Current organizational structures could be used to administer the target, but new organizations should also be considered.
 4. The size of the population should be large and stable enough to avoid large year-to-year fluctuations in expenditures, in order to make the targets predictable.
- o Symmetric incentives should probably be offered (i.e., decrease in unit prices for exceeding the target and increase for meeting it), at least initially, but neither increases nor decreases in price need be equal to the difference between the target and actual expenditures.
- o To increase acceptance of targets by the physician population an appropriate fee schedule should be developed before implementation of an expenditure target.
- o Part A expenditures should be considered in setting the expenditure target since some shifting of costs from Part B to Part A may occur.
- o Mandatory assignment is not necessary but some limit on out-of-pocket expenditures for beneficiaries is desirable.
- o A pilot program with voluntary participation should be instituted to assess the administrative and quality issues raised.

Recommendations: Coding

Because the net effect on volume and expenditures of collapsing procedure codes is unpredictable, we do not recommend it as a Part B volume control. If collapsed coding is considered for implementation at some point in the future, we recommend the following prior to implementation:

- o Patients should be involved in verifying the accuracy of coding.
- o Experimentation is needed with new techniques for coding visits.

Recommendations: Bundling

Bundling should be approached with caution since it could result in distorted incentives for provision of services not included in the bundle. At present, we do not recommend bundling as a volume control. Should it be considered in the future, these factors should be incorporated into its design:

- o Since the idea behind bundling is to reduce the incentive for physician-induced demand, the greater the prospects for demand creation, the more useful it will be to bundle.
- o Services typically provided by the principal physician in a fee-for-service system should be included in the bundle.
- o The more services are substitutable between those performed by the principal physician and those performed by other providers, the more appropriate it will be to include them in the bundle (so as to provide an incentive for the primary physician to provide them in the lowest-cost manner).
- o Conversely, if a service is strongly complementary to the principal service and therefore likely to be provided regardless of whether it is included in the bundle, it need not be included in the bundle since doing so will increase the amount of cost for which the doctor is put at risk.
- o The less variation in severity of disease and in practice patterns that accompany a primary service, the more appropriate it will be to bundle that service with others clinically associated with it.

1.4. Use of Volume Controls by Medicare and Commercial Carriers

All Medicare and most private sector health insurance carriers have utilization review programs, although Medicare carriers appear to focus more on reviewing the appropriateness of physician services than do private carriers. This is partly due to the presence of mandated prepayment screens for Medicare. However, many Medicare carriers had these screens before they

were mandated and, in some cases, their pre-mandate screening parameters were tighter than those currently used.

Several mandated prepayment screens were cited frequently as ineffective, whereas other optional screens were suggested for inclusion in the list of mandated screens. A systematic study of the effectiveness of prepayment screening is warranted, and such a study is currently underway at the HCFA Research Center.

Medicare carriers suggested that new screens be introduced along with programs to educate physicians about these screens. In general, implementation issues are important and should not be ignored. On the other hand, provider resistance to prepayment screens is not necessarily a sign that they are ineffective.

Medicare carriers operate vigorous postpayment review programs. In many cases, physicians are reviewed if their practice patterns are more than 2 standard deviations from the norm. However, the criteria for postpayment review are not uniform. Standards based on reviewing a prespecified number of physicians may be difficult to justify in terms of detecting unusual practice patterns.

Private insurance carriers were found to operate inpatient utilization review programs that combine pre-admission certification, concurrent review, and retrospective review of hospital inpatient care. The respondents estimate that inpatient utilization review reduces total cost by about 7 percent. Many believe that pre-admission certification increases cost and utilization outside the hospital, however.

For managed fee-for-service programs, utilization review has been largely directed at reducing the use of inappropriate inpatient services. PPOs have made far less effort in selecting preferred physicians than hospitals, often using hospital staff privileges as the major screening criterion. HMOs have historically realized their savings by reducing the number of hospital days.

Private carriers have been slow to innovate in the area of physician payment. All but 8 of the firms surveyed by HIAA still pay usual and customary charges. Only 6 respondents utilized "controlled" methods of paying physicians. Commercial insurers appear to place more emphasis on physician payment reform in their PPO arrangements, which frequently use discounted charges or fee schedules to pay physicians. The average discount was estimated to be about 14 percent. This was exceeded in importance by the cost-saving effect of utilization review.

Perhaps the most significant result from our surveys is that neither public nor private carriers have taken an innovative approach toward bundling physician services and collapsing the codes used for paying physicians. The private respondents, by and large, were ignorant of this concept; Medicare carriers generally thought that they had to recognize HCPCS codes for billing. Medicare carriers appear to be ahead of the private sector in using global fees for surgery, but otherwise they have not attempted to bundle physician services into broader reimbursement packages.

Physician education and feedback is utilized more extensively by the Medicare carriers than by private insurers. This may be due to the fact that they lead the private sector in terms of postpayment review, and this is the area where physician education programs are most likely to occur. Medicare carriers, in general, are sensitive to the need for physician education and information programs.

1.5. Volume Effects of the Resource-Based Relative Value Scale

Under a revenue-neutral implementation of RBRVS with mandatory assignment, the price of physicians' services will increase for some services and decline for others, but the change (if any) in the total cost and volume of physicians' services is impossible to predict with certainty. All scenarios rest on untested behavioral assumptions, including the willingness of patients to accept the newly more profitable services and the appropriate model of physician behavior to use to predict physician response.

The most important source of ambiguity comes about because a price change has two conflicting theoretical effects on physician incentives. On the one hand, a physician will want to use fewer of the less profitable services and more of the more profitable ones to treat a condition -- a substitution effect. On the other hand, if the service whose price is cut is an important part of the physician's total business, the price cut will cause income to fall unless he or she can increase demand for other services. If income falls, the physician may create demand for the new lower-priced service to get income back closer to a target level -- an income effect. It is not possible to determine a priori which effect will predominate, and so it is not possible to determine whether aggregate volume will rise or fall with a resource based relative value scale.

2.0. INTRODUCTION

Concern about the volume and intensity of services supplied by physicians has been stimulated primarily by the growth in spending on Medicare's Part B, which, by 1985, had become the nation's third largest domestic spending program (Juba, 1987). With the exception of the years of the Economic Stabilization Program (1971-1974), and the last fiscal year (1987), Medicare spending on physician services has increased at least 10% per year since the inception of the program and increases just prior to 1987 have been dramatic. Between 1978 and 1983, the average annual rate of increase was 20.5% (Burney, et al., 1984). Approximately half of this rapid escalation in expenditures can be attributed to increases in the price of physician services (primarily because of inflation in the general economy, but also as a result of physician fee inflation in excess of general inflation), and a tenth of it is due to an increase in the number of Medicare beneficiaries. However, over two-fifths of the recent growth in spending on physician services is presumed to result from increasing utilization of services on a per capita basis; thus, cost containment within the Medicare program appears to be dependent upon controlling the volume and intensity of physician services provided to beneficiaries.

The implementation of the Prospective Payment System (PPS) within Medicare's Part A has probably exacerbated the concern over physician services. At least some of the savings achieved through the PPS may have resulted from a shift in care from the hospital to the outpatient setting (Fisher, 1987). Given the intent and incentives of PPS, physicians may well be substituting outpatient care for more costly inpatient care.

While the rapid rise in Medicare expenditures has been the primary impetus for concern about physician services, issues of quality and appropriateness of care are also involved. Recent demonstrations of wide variations in physician practice patterns (Wennberg and Gittelsohn, 1982; McPherson, et al., 1982; Connell, et al., 1984; Roos and Roos, 1981; and Chassin, et al., 1987) suggest a lack of consensus on the "appropriate" care of at least some medical conditions. Although relationships between the intensity of care and the outcomes of care have not, for the most part, been established, demonstrations of physicians' modifications of their own practice patterns when presented with information about variations (Wennberg, 1977; Dyck et al., 1977) suggest that there is some amount of flexibility in what physicians themselves consider appropriate practice.

A third reason for concern about the volume and intensity of physician services is that Medicare's current reimbursement mechanism provides varying incentives for the provision of different services. Invasive services are often alleged to be overpriced in relation to evaluation and management services; for instance, four- to five-fold differences in payment per unit time for inpatient surgery and office visits, after adjustment for complexity, have been documented (Hsaio, 1979). Physicians thus make larger profits if they substitute invasive services for evaluation and management services when possible. To the extent that such substitution occurs inappropriately, the quality of care is adversely affected and costs may be increased. These differing incentives may also have long term effects on the distribution of

health manpower (both geographically and in terms of physician specialty) and on the development of new technologies.

Finally, the significant increase in volume and intensity of physician services and the resulting impact on cost forces higher premiums. The 1988 increase of 38% in part B premiums is a case in point. This increase reduces what Medicare beneficiaries can spend on other things.

2.1. Factors Affecting Volume and Intensity

A number of factors are thought to influence the volume and intensity of services provided by physicians; these include the financial incentives presented to the physician by the payment mechanism, the physician's value system, the characteristics and desires of the patient, the characteristics of the broader system within which care is delivered, the technology available to treat illness, and administrative or financial controls implemented explicitly to restrain volume and/or intensity. The relative influence of each of these (or other) factors on physician behavior is not known, but consideration of their qualitative effects, and empirical study of their quantitative effects, should precede attempts to control the volume and intensity of services.

Much of the HCFA cost-containment activity of recent years has been based on the premise that payment of physicians on a fee-for-service (FFS) basis at current levels is more "inflationary" than capitation, in that FFS payment offers financial incentives to provide a greater number of services or more expensive services. This hypothesis, in turn, is based on the assumption that physicians respond, at least somewhat, to financial incentives. Such a behavioral response to FFS payment may take several forms: providing a higher number of services, substituting more profitable services for less profitable ones, providing services at sites where reimbursement is greater, "bundling or unbundling" services to maximize reimbursement, or restricting patient access to services with negative profits (Eisenberg, et al., 1987). In contrast, capitation theoretically provides financial incentives to limit the number of services, to consider cost (rather than profit) in choosing the mixed sites for services, and to limit access to all services.

Empirical evidence for physicians' ability and willingness to "create demand" for more services or for certain services is not conclusive. During the national Economic Stabilization Program, in which payments were frozen, Medicare services increased (Gabel and Rice, 1985; Yett, et al., 1983; Eisenberg, et al., 1987), while the volume of private services per physician fell (Yett, et al., 1983). However, at least some of the increase in Medicare services was attributed to "unbundling" of services (Eisenberg, et al., 1987). When Medicare payment rates were decreased in Colorado, increases in medical, surgical and laboratory services were observed (Rice and McCall, 1982). Some evidence exists that more lucrative services may be substituted for less lucrative services when relative payments change (Marquis, 1982; Rice and McCall, 1982; Danzon, et al., 1984; Danzon, et al., 1980; Danzon, et al. 1983; Eisenberg, et al., 1987).

On the other hand, direct tests of demand creation on the quantity of services at a given price have often (though not always) found effects that are small (Wilensky and Rossiter, 1983; Pauly, 1980), or explainable by changes in quality or declines in non-monetary prices. Phelps (1986) has recently expressed skepticism about the ability of empirical data to determine whether "true" demand creation occurs. But there have been relatively few studies, and no definitive studies, directed at changes over time in volume and intensity as a function of changes in the price or availability of physician services. What is perhaps equally important, evidence that price decreases cause increases in volume is not consistent with the expectation (evident the rationale for such changes as the RBRVS) that price decreases for, say, invasive services will cause decreases in the volume of those services.

That is, it is unclear whether services that are made more lucrative by a change in relative prices would always be provided in greater amounts, especially if those services are concentrated in and a substantial part of the income of particular specialties. A scenario can be imagined in which physicians who typically provide the newly profitable services will find their incomes increasing if they simply maintain previous levels of provision, while physicians who provide the now less lucrative services would need to increase volume in an attempt to maintain their previous income. A change in relative prices might thereby stimulate provision of the services whose profitability has been reduced and dampen provision of the services whose relative prices were increased. An alternative scenario could also be imagined, however, in which short-term changes in physicians' practices led to substitution of the services with an increase in price for services with a decrease in price, and in which long term changes in physician specialty or geographical distribution caused a similar result.

The financial consequence of a practice decision is not the only factor influencing that decision, however. Physicians have as their primary goal the improvement of patients' health (Eisenberg, et al., 1987) and are legally and ethically bound to "do no harm". Thus, the response to financial incentives is constrained by these other concerns. The extent to which physicians do respond to financial considerations is likely to depend largely on their beliefs about the patterns of care that represent high-quality (or even adequate) care: if a range of practices are considered equally appropriate in a particular instance, physicians may be able to respond to financial incentives by providing the most profitable, rather than the least profitable, services while still providing appropriate care; whereas, if appropriate options are more limited, the physician will be less able to respond to financial incentives. The values that individual physicians place on providing high quality care versus the value placed on maximizing or maintaining income will influence their behavior in this regard.

Certain characteristics of the patient will also influence the services provided by the physician. Patients may be unwilling to accept services without limit, and a demand-inducing physician may suffer a loss in reputation (Pauly, 1980; Dranove, 1984). A medical/sociological model of the determinants of utilization of ambulatory care developed by Anderson and Newman (1973) identifies three relevant types of variables: predisposing

(socioeconomic and attitudinal factors that encourage or discourage the use of ambulatory services); enabling (income, insurance status, and supply of physicians); and health status (indicators of the existence or severity of perceived or diagnosed conditions) (Buczko, 1986). In a number of studies, the health status indicators were found to be the most important determinants of utilization, suggesting that patients' signs, symptoms and apparent severity of illness influence physicians' choices to a significant degree.

Finally, certain characteristics of the system in which care is delivered may affect the volume and intensity of services. Defensive medicine attributable to the threat of malpractice liability, for instance, is thought to stimulate the use of diagnostic tests and procedures. Quality assurance and utilization review activities may encourage or discourage the use of certain services. In addition, the kind of technology available to a physician will influence his or her provision of services.

Available empirical information is therefore inconclusive, and policy may need to be decided before additional research can provide more definitive information. This report examines the rationale for and potential usefulness of a number of schemes which have been suggested as methods of controlling volume and intensity while maintaining quality and access at acceptable levels. The discussion is explicitly based on an assumption of imperfect information. That is, we discuss each scheme, and the possible forms it can take, with explicit recognition that a number of the critical questions about the behavior of doctors and patients do not yet have firm answers. Consequently, schemes which will produce good outcomes under a wide variety of possible behaviors will be considered preferable to those that may produce the best results under one scenario but have the potential for seriously adverse results in other, plausible circumstances. In effect, we will be commenting on whether a strategy has "can't lose" properties, as well as on its performance under ideal circumstances.

This report focuses on the volume and intensity consequences of alternative schemes. For many of these schemes, quality and (to a lesser extent) access questions have been investigated by others (Hammons, Brook and Newhouse, 1986), so we provide less detailed consideration of those consequences. We do, however, pay attention to quality and access dimensions, especially in the case of schemes which have yet to be subject to detailed analysis elsewhere.

3.0. MODELS OF PHYSICIAN BEHAVIOR

Physicians are paid for their services in part to offer them incentives to furnish those services. And yet there are some reasons--so far not fully conclusive--to be doubtful that the volume of those services is ideal. There are two broad policy options. One could change prices or payment mechanisms in some way: these changes must, at some point and in some fashion, affect volume. Alternatively, one could impose non-price controls which overrule the decisions on which doctors and patients would otherwise agree. But obviously neither Medicare nor any other insurer can review and dictate the process of care in every patient-physician encounter. So one must develop some predictions, based in part on experience with prices and controls and in part on other bases for understanding how doctors and patients behave to forecast the effects of various "price" and "non-price" volume-affecting schemes.

An essential element in thinking clearly enough to make predictions with some degree of confidence is an idea of what physicians are trying to do. What objectives are they trying to pursue when they produce and sell services to patients? Here we describe three alternative concepts (or models) of physician behavior. In Appendix I, a broader consideration of financial incentives in medical decision making is presented.

3.1. Profit Maximization

3.1.1. Description of Model

The simplest story is that doctors are like other small businessmen. They try to maximize profits, at least within the limits of the effort that human beings are willing to make. This does not preclude ethical behavior or concern for patients, especially when ethical behavior becomes known and reinforces demand. But it does imply that, at least as a caricature, decisions are made largely in terms of their effect on the long run bottom line. Even if one does not believe that physicians actually behave this way, it does provide a benchmark, a pure case of what someone with solely financial objectives might do, and therefore may serve as a stylized predictor of what financial incentives will do.

In order to discuss outcomes in a model in which physicians are only interested in profits and in which demand inducement is possible, we obviously have to characterize the reaction that patient/customers might have to inducement. After all, if there is no reaction, the amount of demand creation for a profitable service would be infinite! In order to emphasize the difference between assuming profit maximization or "target income" as an objective, we assume that the services for which doctors are paid have independent demands, and that there is a maximum amount patients will accept of each service. That is, we imagine that the services are used to treat different illnesses (even though some are provided by the same doctor), and that there is a maximum amount of the service that a patient/consumer will accept for a particular type of illness. The situation in which services are regarded by patients (and doctors) as substitutes or complements in the treatment of a particular illness will be considered in a later section of

this report. The qualitative conclusions for that case will be the same as those we derive here, but the argument is considerably more complex.

3.1.2. Profit Maximization and the Effect of Price on Volume

Profit is the difference between revenue and cost. For an owner-managed firm, such as a medical practice, "cost" includes not only explicit cash payments for inputs made to others but also the money value of the owner's time which he or she sacrifices by working in the business. The simplest strategy here is just to price out owner-provided time at some valuation per hour, in much the fashion that the Urban Institute researchers did in constructing a geographic economic index of physician practice costs.

When nominal output prices are frozen and real output prices fall in an inflationary environment, how would such a profit maximizer respond? The answer is straightforward--with no objectives other than profit, the firm would decide to sell less output. The reason is obvious. Before the price cut, the firm would have been willing to sell output up to the point at which the last unit's cost was just covered by the price received from Medicare. But if Medicare cuts its payment price, this last unit becomes a money-loser, and so it (and enough other units in the same situation) will not be sold.

There is a critical assumption behind this common sense conclusion. The price cut has to force price for some services below cost. If the initial price was substantially above cost, and the price cut just pushes it closer (but not below), then the common sense conclusion will not follow. A firm interested only in profit will continue to provide a service even if it becomes less profitable, as long as there is still some money to be made. The real question then is how many services will a particular price reduction shift from "gainers" to "losers". There probably are always some such services provided by some physicians just at the margin and presumably these physicians will be unwilling to continue treating Medicare patients in the same way when price falls. The empirical issue is how thick this margin is. What is unequivocal is that price cuts to a profit-motivated supplier, will either reduce volume or, in the limit, have almost no effect on volume. In this scenario, price cuts would certainly not lead to an increase in the volume of services provided.

3.1.3. Adding Inducement

But, one might object, wouldn't one expect profit seeking sellers to try to recoup some of the lost total profit on services whose profit, though still positive, has been reduced by a price cut? The surprising answer to this question is "No". If the physician was and is only interested in profit, if he or she always makes profit as large as it can be, and if there is a possibility of demand creation; then Profit Maximization Theory implies that, even prior to the price change, all profitable demand creation will have been created. Consequently, there will be no profitable niches which can be tapped. Put slightly differently, profit maximization with demand creation implies exploitation of all profitable opportunities before the real price is

cut. Since the price cut does not create any new opportunities for profit, there is nothing left to be exploited. In this caricature, all of the gains from Medicare and the consumer have been wrung out of the system.

While there is a rich set of further implications of profit maximizing behavior that we pursue elsewhere in this report, this conclusion is sufficient to suggest that we need to develop another story, one in which doctors have other, more noble or complex objectives than profit (such as practicing the highest quality medicine).

3.2. The Sophisticated Target Income Model

If doctors are interested in more than maximizing the net income that they get from Medicare practice, given the time they put into medical practice, one popular alternative formulation is to suppose that they have a target income, rather than the maximum possible income, in mind. The simple version of this story imagines that there is some fixed number of dollars, given time and effort, to which a doctor aspires. While this simple approach is useful as a first approximation, it has some obvious logical problems. One problem is that this story implies that the doctor will do anything to make more money if he or she expects income to be below the target, but is not attracted at all by the prospect of extra money once income hits the target. Nature, including human nature, is unlikely to make such a leap in valuation. Put slightly differently, it seems implausible to imagine that a person would do nothing to get income above the targets and anything if income is below the target. Something must constrain the person from going all the way to income maximization.

The other logical problem is that the single target income theory is incomplete; it is incapable of answering certain crucial questions. For instance, suppose a doctor is forced to accept a lower price for one of the services he or she sells. The target income story says that volume will be increased to make up for the lost income, but will the increase fall on the service whose price was cut, on all other services, or on some other specific services which are highly profitable? The simple target income story just cannot give an answer.

3.2.1. The Sophisticated Model

To solve both of these problems, analysts have been developing a more sophisticated version of a target income story. This version solves the problems just discussed by imagining that the doctor does give up something when he takes actions to raise income (given time worked and inputs hired). The notion is that there is some set of services which the doctor knows are most appropriate for patients, and which patients would demand if they were given accurate, complete, and perfectly truthful advice. (The advice depends, of course, on what the doctor knows, which is usually much less than is needed for absolute certainty). The doctor will receive some positive income if this advice is given, or this ideal quantity provided, but that income may be increased (moved closer to a target) if the doctor "induces" or "creates" more

demand beyond the "full truth" point. As we showed earlier, if only money matters to the doctor, he or she will always induce the maximum amount of this demand. But it may be plausible to assume that the doctor suffers a kind of psychic cost from deviation from his or her heart-of-hearts belief in the best form of therapy or the best level of advice. There is, after all, sufficient uncertainty and ambiguity in knowledge about the effectiveness of care to justify more service-intensive types of care, especially if these services do not put the patient at much more risk and only take the patient's time or Medicare's money.

So the sophisticated target income model assumes that there is a kind of "cost of conscience" to creating demand, a cost that at least some doctors are sometimes willing to pay to enhance income a little, but which ultimately limits their willingness to induce demand. Some doctors probably increase their income by providing a more intensive style of care than they would feel perfectly comfortable with, but which they judge to be not unreasonable, given the reward.

The preceding paragraph argues that the physician would be inducing demand only if he or she creates demand beyond the "full truth" point. But consider the situation in which the patient desires a level of services that is below what the physician believes would maximize health so that there is a gap between what the patient believes he or she needs and what the doctor would provide if he or she were allowed to (independent of any effects on income). At present, assume that the doctor may not be providing additional information in order to convince the patient that more care would be beneficial, because the time and effort required to convince the patient would be more costly to the physician than the additional revenue would justify. However, if the physician finds that there is now an economic reason for changing the amount of information provided to the patient (e.g., an increase in price or a decrease in the resources required to convince the patient), then the physician might provide information to the patient in hopes of narrowing the gap between what the patient desires and what the physician believes would be beneficial to the patient.

Another consideration is that the doctor may realize no psychic cost for increasing the amount of services recommended to the patient if, in some subconscious way, the higher price for a service changes the way in which the physician interprets the literature about services for which indications are uncertain or ambiguous. This implies that the "heart-of hearts belief in the best form of therapy or the best level of advice" is not fixed and might be influenced by messages provided to the physician from the payment scheme. It could even be argued that society's message to the physician about what it values is implicit in the price that it is willing to pay for services. Behavioral responses to these price messages by physicians would be consistent with the Sophisticated Target Income Model.

3.2.2. Determinants of Inducement

The critical idea then is that changes in the level of volume brought about by changes in the level of inducement can be best understood by looking

at changes in the terms of tradeoff between money income and psychic comfort. For instance, if the price of one service is raised a great deal, this story says that the doctor will be sorely tempted to induce more demand, since the psychic cost will be the same as before but the reward will be greater. But the theory also says that the doctor will not overdo it on recommending more profitable services, because his or her ethical sense will (eventually) be offended by trying to encourage use by patients who really do not stand to get much benefit from the service.

The final critical idea for this model notes that the change in the tradeoff really depends on two things. It depends on the "marginal net income reward" just discussed. But it also depends on the value the doctor attaches to being able to practice ideal medicine. And the notion is that this value may itself be affected by the doctor's income. When his or her income is low, with heavy obligations and high fixed costs, less value may be put on practicing medicine in an ideal way than if a windfall should dramatically raise income. This notion helps to build in the possibility of target income behavior. Specifically, suppose the price of all services a doctor provides to Medicare patients were to be cut. Then income would fall, and with it the reluctance to induce demand. So we might not be surprised to see additional inducement of demand for all services (not necessarily in a uniform pattern, however), even though the profit from doing so is less, because the doctor feels an increased need for the money.

3.3. The Patient Agency Model

While much of the relevant economics literature emphasizes physicians' responsiveness to changes in the price paid for their services, other models of behavior do not rely on direct economic explanations of physician behavior. One of the most popular non-economic models of physician behavior suggests that they seek to serve as their patients' agents. A substantial portion of the physician's satisfaction with practice is fulfilled by serving successfully as the patient's advocate. In this role, the physician makes decisions that represent what is in the patient's best interest -- at least what the physician perceives to be in the patient's best interest. Thus, the physician as the patient's agent makes decisions that he or she believes are the decisions the patient would make were the patient to have as much information as the physician has. As with profit maximization, this perfect agency model serves as a benchmark for purposes of predicting effects on volume and intensity.

We suggest that there are several components to this model of medical decision making. First, the physician's primary role as healer demands that he or she attempt to optimize the patient's clinical outcomes. Second, the physician will also want to serve as the patient's economic agent, trying to make decisions about the use of medical services with their financial impact on the patient in mind. Third, doctors will be influenced by their patients' health preferences, which are manifest through patient demand for medical care. Finally, patient convenience and other preferences of the patient will influence the decisions of physicians who are attempting to act in their patients' best interests. This model is described in chapter 3 of

Dr. Eisenberg's book, Doctors' Decisions and the Cost of Medical Care (pp. 57-77).

3.3.1. The Clinical Agent

Physicians' concern for their patients' health has been shown to be a major determinant of utilization patterns. Physicians' principal professional motivation is to apply their knowledge and ability to improving or maintaining the health of individuals who have turned to them for help. The unifying hypothesis is that physicians have personal utility for their patients' health and that this utility overrides the physician's concern for his or her own financial welfare.

Clinicians, such as Wenberg, writing about differences in physician utilization assume that a doctor would want to give better care if it were possible. However, there is substantial uncertainty involved in medical decision making, and the strategy that would optimize patient health is not always clear. This uncertainty may exist because data are not available or because the available data cannot be processed effectively. It is also possible that data are available but not known to the individual physician making the decision. The reasons for deficiencies in a practicing physician's knowledge base are many, but the psychic and financial cost of acquiring new information are two important factors influencing the availability of data to reduce uncertainty.

Given this uncertainty in any particular patient-doctor encounter, it is not surprising that physicians seeking to provide optimal care are influenced by forces outside the immediate doctor-patient interaction. In particular, the doctor's knowledge or opinion of what constitutes best care will influence his or her action. When there is uncertainty about indications, procedures, or protocols; variations in decisions about what constitutes best care are likely. Given this scenario, physicians' decisions can be changed by providing them with better, or more authoritative, advice about appropriate treatment methods. Changing the price the doctor gets, in this model, will not influence variations in practice patterns, but changing information on standards will.

3.3.2. The Economic Agent

In addition to the influence of objective information, and the recommendations and advice of peers and professional leaders, the perfect physician agent will be influenced by financial considerations from the perspective of their impact on the patient's well-being. Evidence does indeed suggest that physicians are sensitive to the financial impact of medical care on their patients, even though this effect may be smaller than those concerned with cost containment would desire. Despite their generally poor knowledge of medical care prices, patients' insurance coverage and patients' financial status, physicians do seem to respond to the cost of care to their patients. Some evidence suggests that the elasticity of demand for medical care may be traceable more to physicians' decisions on behalf of patients than to

patients' own decisions (e.g., patient versus physician-initiated visits, use of laboratory tests, hospitalization).

3.3.3. Patient Preferences and Demand

Patients' demand for medical services is a manifestation of several factors, including user price, convenience, perceived value and alternative sources of relief. As the patient's agent, the physician will want to consider these factors in choosing the appropriate clinical strategy. Perhaps most difficult for the physician is understanding the patient's preferences, or utilities, for various health states. Since patients may find it difficult to articulate their preferences, and physicians may find it difficult to discuss their patients' values, physician decisions on behalf of patients may be misguided. If a physician attempts to serve as the patient's agent but does not understand the patient's utilities, it is possible that the patient's preference will be misrepresented during the physician's decision making process. If better information about patient values, and costs, could be communicated to the doctor-as-agent, decisions about volume and intensity would often change.

3.3.4. Patient Convenience

The convenience of the patient also may influence utilization patterns in accord with the agency model. For example, patients may prefer to have laboratory tests drawn by their own physician rather than going elsewhere for venipuncture. Economic factors inevitably interact with the desire for convenience. While patients who travel farther for laboratory testing may find lower prices, they also incur greater travel costs and greater opportunity costs in doing so.

3.3.5. Other Factors

The behavioral elements of medical decision-making discussed thus far consider only the role of the physician as the patient's agent. Another set of behavioral factors has to do with the physician's own role in the profession and his or her personal preferences. The powerful influence of professional leadership offers another behavioral model with profound implications for peer review and educational influence (Eisenberg, 1986). Physicians will also be influenced by their desire to employ a particular styles of practice, their own personal characteristics, and the practice setting. The physician's desire to serve the social good is another non-economic behavioral model that needs to be considered in predicting or understanding physicians' responses to changes in the medical care system, particularly those designed to influence their behavior.

4.0. DESCRIPTION OF CONTROLS

4.1. Clinical Guidelines With or Without Education

Guidelines are defined as information which delineates specific criteria for the use of a particular test or treatment. Guidelines specify "appropriate" treatment; once disseminated there is typically no monitoring or enforcement. This information is usually disseminated in written form, but may be communicated or reinforced orally. One might think of guidelines as universally and prospectively disseminated criteria. The criteria may be one or more of the following types:

- a. Physiologic criteria consist of specific values or cutoffs for physiologic measurements. For example, "O₂ therapy is indicated when PaO₂ is less than 60 torr."
- b. Clinical indications usually involve observable symptoms of some type. For example, "wheezing with secretions" may indicate the need for a specific respiratory therapy treatment. Indications might also be related to clinical history, such as "recent pulmonary surgery."
- c. Risk factors refer to facts in the patient's history that are significant enough to indicate the use of a test or treatment. These may be quantifiable, such as ">40 pack-years of cigarette smoking," or not quantifiable.

Guidelines can be formulated in several ways:

- a. As algorithms or heuristics
- b. In probabilistic form
- c. In text-book or pocket guide form

The face validity of the guidelines is also important. Relevant factors related to the face validity are from whom they come (NIH versus a drug or equipment salesman), what purpose or audience they attempt to address, and how they were developed (e.g., the data base used). Guidelines have frequently been studied in conjunction with other methods of influencing practice patterns, such as education, audit, and feedback.

4.2. Utilization Review With or Without Penalties

Utilization review is review of the patient's medical record and/or claim through application of defined criteria. These criteria are often derived statistically i.e., by looking at the average practice patterns of a large population. The purpose of utilization review is to assess the efficiency of the health care process and the appropriateness of decisions regarding the site frequency, and duration of care. Inpatient utilization review can be conducted either before admission (pre-admission review), during the patient's stay (concurrent review), or after discharge (retrospective review).

The method of utilization review can be implicit criteria, explicit criteria, or both. Utilization review is an important component of utilization management, which is deliberate action by payers to influence hospitals, physicians, or other providers to increase the efficiency with which hospital services are provided. Utilization management is often conducted by a(n) (external) case manager. Such managers are likely to perform utilization review prior to determining a case management strategy.

Interventions may be made by the payer, an agent of the payer (such as a utilization management vendor), or a provider to reduce inappropriate utilization. Interventions include the following:

- a. Providing feedback to physicians e.g., comparing a physician to a group of similar physicians treating similar cases;
- b. Conducting education, which may include practice protocols;¹
- c. Restricting benefits, such as not allowing surgery without a second opinion and;
- d. Imposing sanctions or penalties, such as withholding payments from physicians or restricting hospital privileges.

4.3. Copayment

Copayment refers to cost-sharing, by the beneficiary, of a portion of his or her medical bills for covered services. For the purpose of this report, two forms of copayment are defined: deductibles and coinsurance. A deductible, defined in dollars over a specified time period, is a set amount of out-of-pocket expenditure a patient must incur for covered services before his/her insurance begins to pay for the services. Coinsurance, defined as a percentage, is the proportion of expenses for the covered services that the patient is responsible for after the deductible is met.

4.4. Capitation With or Without Financial Incentives

Capitated payment, as the name implies, is a fixed "per-head", per-period payment to a health care provider for a defined set of benefits, regardless of actual utilization. This payment method essentially merges the insurance function with the provider function, that is, the risk is borne by the provider. The "provider" may be a health maintenance organization, hospital, physician, physicians' group, or other provider.

¹Utilization review using explicit criteria with education and feedback is very similar to guidelines with education. These can be distinguished by viewing guidelines as universally and prospectively disseminated criteria, while utilization review with feedback is retrospective and more likely to be targeted at outlier physicians.

The key question in a capitation arrangement is "Who is being capitated for what?" The provider whose services are being capitated and the comprehensiveness of the benefits for which provider is responsible are central to the incentives the arrangement presents.

When the capitation is at the HMO level, the incentives to the physician vary with the organizational arrangements. HMOs receive capitated payments for providing a comprehensive set of benefits to subscribers. The impact of an HMO on physician behavior will depend on the following factors:

- a. The organizational model of the HMO
 - staff model
 - group model
 - network model
 - IPA model
- b. The method of paying the physicians
 - fee-for-service
 - salary
 - capitation
- c. Additional financial incentives
 - over-all financial performance bonus/penalty
 - individual-specific performance bonus/penalty

Most research evaluating the impact of capitation on utilization rates has focused at the HMO level. In the HMO, the attending physician is a (internal) case manager. Little research has been done looking specifically at the impact of capitation payments to physicians on utilization. The strength of the utilization and cost control incentives facing the physician will depend on the comprehensiveness of the services his capitated payment covers and on the specific capitation arrangement.

4.5. Expenditure Target

An expenditure target is a global budget covering a defined population of beneficiaries. There are three elements of an expenditure target as a physician payment mechanism:

- a. An expenditure projection model to set the global budget or expenditure limit.
- b. A unit price adjustment formula. Unit prices are adjusted downward if total expenditures exceed the budget (and possibly adjusted upward if expenditures are under the budget).

- c. The timing of the price adjustment. Unit prices can be adjusted in the current year (e.g., through a partial withholding of payments) or in the following year.

There has been little or no experience in using this method to pay physicians in the United States.

4.6. Collapsed Procedure Packages (Coding Changes)²

Over time, CPT-4 coding terminology has become highly fragmented with many codes for a single basic procedure. It has been proposed that as the number of procedures grows larger, and the distinctions between them blur, procedure inflation results. Procedure inflation is the practice of billing under a more complex (i.e., more expensive) procedure code. The idea is that the more numerous the codes for a service, the easier it is to for instance, subjectively rename a brief visit an intermediate one, or an intermediate visit an extended one.

Collapsing procedure packages is intended to reduce or reverse procedure inflation by collapsing similar procedures into one code in a way that retains fundamental procedural distinctions. Collapsed procedure packages have been considered for surgical procedures and for office visits. For example, the eleven different office visit codes that currently exist could be collapsed into as few as two different payment codes.

4.7. Service Packaging (Bundling Services)

Un-packaging is the practice of submitting an itemized bill for every service performed. Service packaging is the combining of services related by diagnosis or procedure into one reimbursable service. The idea is similar in form and incentives that are used in Medicare's prospective payment of hospitals by DRG's. Three types of packaging have been proposed:

- a. Office Visit Packages: An office visit packaging arrangement would base reimbursement on a per visit basis and would include all associated ancillary services. Numerous problems have been pointed out with this approach, most notably the within-package variation in the use of services.
- b. Ambulatory Condition Package: Under an ambulatory condition package the physician would be responsible for all aspects of a patient's treatment for a predetermined period of time. A lump-sum payment would be made to the physician regardless of the type or quantity of care provided. Covered services would include all visits to the package physician, all

²The definitions of collapsed procedure packages and service package bundling follow the discussion in: Mitchell, Janet B., et. al., "Packaging Physician Services: Alternative Approaches to Medicare Part B Reimbursement;" Inquiry 24: 324-343 (Winter 1987).

outpatient diagnostic and therapeutic services, and any consultations with other physicians. A major problem with this kind of package is the extent of out-of-package or out-of-condition care.

- c. Special Procedure Package: Special procedure packaging combines all services directly related to a procedure into a single bill and makes a lump-sum payment to the physician responsible for the procedure. This type of packaging arrangement could be used for all surgical operations, all invasive diagnostic tests (such as endoscopies and cardiac catheterization), and complex radiological procedures. It would be appropriate for all locations -- office, hospital, ambulatory surgical center, and so on. Special procedure packages would generally, but not always, bundle together the services of multiple physicians.

5.0. ANALYSIS OF CONTROLS

Before discussing the anticipated effects of these various controls, some issues that apply to volume control in general should be noted. First, for any attempt to control Part B volume, an explicit decision would be needed about whether the control would be mandatory for all physicians and all beneficiaries, or whether certain groups could "opt out" in return for lower unit prices or some other concession. A determination concerning whether to allow balance billing would need to be a part of this decision. As at present, it may be reasonable to allow those who choose to pay for additional services with their own funds to do so, but the implications for the particular control and its overall effectiveness should be considered beforehand.

There is a tradeoff here. Using mandatory assignment would induce or compel some doctors to accept the lower Medicare price, and restrictions on balance billing, rather than forego Medicare patients. The beneficiaries who use these doctors, and the Medicare program, would presumably be better off. But some doctors may refuse to accept assignment; beneficiaries who would have used them are made worse off, the more so because the beneficiary is not even permitted to volunteer to pay to supplement what Medicare pays in order to get a preferred doctor. Some beneficiaries will benefit, others will be harmed. It is difficult to say whether the average beneficiary will win or lose, or even whether the well-being of the average is a relevant measure of desirability. The answer may depend in part on whether the higher prices mandatory assignment prohibits are thought to yield monopoly returns or not.

Secondly, it should be noted that the controls discussed here need not be implemented in isolation; rather, combinations may be more effective. Clinical guidelines, for instance, would be a useful adjunct to any other volume control. Likewise, rigorous Part B utilization review could be implemented along with an expenditure cap or with changes in copayments so as to monitor the appropriateness of services provided. One of the major reasons for considering collapsed procedure coding as a control is that it would allow for closer monitoring of upcoding; in this case, expanded utilization review would be necessary for effectiveness. Other techniques could be used with the controls described here as well: case management, for instance, could readily be used in conjunction with utilization review; and capitation, with the current or a modified version of copayment, or under an expenditure cap. Thus, while seven controls are analyzed separately here, these possible combinations should be kept in mind.

Thirdly, effects of any Part B volume control on Part A volume and expenditures is key if there is to be stabilization or reduction of overall program costs. There is little sense to an approach that merely shifts costs from Part B to Part A and, while this is somewhat constrained by the prospective payment system, some degree of shifting is possible. Effort should be taken to avoid inappropriate changes in the locus of care and attendant increases in overall program costs.

Long term effects on health manpower are also a consideration. The degree to which volume controls discourage entry into the medical profession

in general or into individual specialties is important to the broad objectives of the Medicare program and needs to be examined.

Finally, we would recommend that the widespread implementation of any of these controls be preceded by a demonstration project. Our predictions of what is likely to happen with each control are based on thoughts about how physicians as a group and, to a lesser extent, patients respond, but very little empirical evidence exists to substantiate these predictions. Further, as pointed out in this report, a variety of design issues for each control bear on its ultimate effects; the final form that a control takes will therefore influence these effects. A demonstration project would be needed to predict these accurately and avoid or minimize adverse consequences.

5.1. Clinical Guidelines With Professional Education

5.1.1. Overview of the Literature

Reports of the effectiveness of clinical guidelines in influencing medical practice have been mixed. To some degree, this is undoubtedly because there have been a variety of types of guidelines and methods of dissemination studied. Guidelines can be developed for two purposes: (1) to inform, educate, and guide providers (especially physicians); and (2) to control service utilization. These different types of guidelines can then be disseminated in various ways, for instance, with or without education, with or without penalties, with or without the support of clinical leaders, etc. Review of the literature on the effectiveness of guidelines is therefore complicated by these variations. It is also of note that much of the empirical work in this area has been done in the inpatient setting rather than in ambulatory practice; application of these results to Medicare's Part B should thus be considered carefully.

Perhaps the most compelling work supporting the use of guidelines comes from Wennberg and Hanley, in collaboration with the Maine Medical Association (Wennberg, 1985; Physician Payment Review Commission, 1988). While this effort is still underway and has not yet been formally evaluated, early experience suggests that guidelines as simple as feedback showing one physician (using a small set of physicians in a given specialty) his or her own practice in relation to peers can be effective in altering patterns that deviate from the statistical norm. In addition to this direct influence on practice, such an approach has been reported to stimulate research on the relationship between various rates of service utilization and outcomes of care (Wennberg, 1985).

In his review of the effectiveness of various methods of influencing physician behavior, Eisenberg (1985) found that such feedback is particularly useful as an adjunct to educational guidelines. The most effective feedback appears to be that provided in a personal manner by a respected professional, individualized to the physician and representing current or recent practice data. Such feedback can be expensive, however; in a study of physician education with chart audit and feedback by Schroeder, et al. (1984), the

relatively small financial savings were thought to be negated by the costs of the program.

The effect of educational guidelines (i.e. those that are based on outcomes research or consensus of expert opinion and that suggest clinically appropriate care) appears to depend substantially on the method of dissemination (Eisenberg, 1985). Although most studies suggest that the simple transfer of information is not effective in changing practice, certain factors appear to enhance effectiveness. These include the use of computer reminders, the commitment of clinical leaders, personal as opposed to impersonal transfer of information, and an environment in which physicians are open to change. A recent study of the effect on physician practice of the recommendations of the Consensus Development Program of the National Institutes of Health supports this conclusion: in this situation, where results are simply published in a leading professional journal, little effect was seen on practice (Kosecoff, et al., 1987). Similarly, one recent survey of the practice patterns of internists has demonstrated relatively low levels of compliance with the cancer screening guidelines of the American Cancer Society and with the immunization guidelines of the Centers for Disease Control (American College of Physicians, unpublished data).

Although capitation has the potential to control Part B volume and expenditures, a number of problems related to capitation make it less attractive as a volume control than other approaches. Organizational changes necessary for a completely capitated system would be difficult to realize in the short term. Capitation on a partial basis raises the possibility of increased referrals for services not included in the partial capitation, which could preclude significant cost savings. Either partial or complete capitation present risk pooling problems that require case mix severity adjustments; at present, these methods are not well developed. Likewise, either type of capitation requires an equitable determination of capitation fees. Given the current problems with AAPCC's for Medicare HMOs, a successful approach to setting appropriate and equitable fees seems unlikely in the near future.

5.1.2. Likely Effects of Implementation

As noted above, clinical guidelines might be developed for two purposes: (1) to inform, educate, and guide providers (especially physicians); and (2) to control service utilization. While the two objectives are not mutually exclusive, they do have different implications regarding guideline development and implementation.

5.1.2.1. Background

Modern medicine attempts to be scientifically based. Ideally, the provision of medical services is based upon considerations of safety, efficacy, effectiveness, benefits, and costs. This information is obtained through clinical observation and scientific study.

There is both direct and indirect evidence that medical practices are not always used appropriately. Studies of specific technologies and of groups of patients with specific diagnoses have often identified misutilization of specific services when observed practices are compared to explicit expert guidelines. Studies of services across both large and small areas have identified large variations in utilization. Although the reasons for such observed variations have not been well studied and are not well understood, most experts believe that a substantial portion of the variations cannot be explained by differences in medical need or patient preferences. Rather, these variations reflect differences in provider assessment of the value of the service.

It is not possible for physicians to evaluate independently the impact of the great volume and range of available clinical practices. The volume of the medical literature is huge, of variable quality, widely diffused among a large number of journals in a variety of specialties, and often not clinically oriented. Substantial uncertainty surrounds many clinical practices and procedures. Moreover, the demands of clinical practice operate against individual decision making in many common and routine patient encounters. Thus, much of medical care is based upon physicians' heuristics and habits.

Development of expert guidelines regarding the appropriateness of specific services for specific groups of patients offers the opportunity to provide physicians with improved information for the care of their patients. When such recommendations are carefully developed through a process in which providers have confidence, they are widely accepted (e.g., the CDC's ACIP immunization guidelines). Under appropriate circumstances (discussed below), such guidelines can result in more informed providers, more appropriate service provision to patients, improved patient care and outcomes, and, often, reduced service volume and costs.

The primary appeal of such guidelines is that they approach utilization of services primarily from the perspective of the patient and the public. Clinical practice guidelines focus discussion on issues of efficacy, effectiveness, and safety, thereby narrowing the range of conflict.

5.1.2.2. Design Issues

Guideline development and use have two types of components: (1) medical-scientific components; and (2) public policy components. Medical-scientific components refer to objective determination of the safety, efficacy, effectiveness, benefits, and costs of specific services under specific clinical conditions. These components establish the facts regarding the clinical impact of services. Policy issues include comparing the tradeoffs of the medical service with those of other medical and non-medical services, and evaluating the impact of the service on costs. Society or the payor judges if the benefits are worthwhile given the costs. This involves ethical and political concerns and reimbursement decisions.

Guidelines could be used to guide coverage policy per se; that is, recommendations on safety and efficacy can dictate whether a procedure or

service is covered. Once a service is covered, utilization can be used to meet a variety of different objectives. If guidelines are used to deny payment for applications not covered or for inappropriate use of the service (i.e., indications for appropriate use are not met), they can be thought of as establishing boundaries on practice. However, if they are used to encourage ideal medical practice, they are more accurately thought of as establishing pathways, and allow considerably more leeway for physician decision making. Both in formulating and implementing guidelines, the objectives sought should be clear.

Physicians and patients both perceive physicians as having the greatest expertise regarding issues of patient care and as being the party most likely to protect patient interests. Therefore, the success of a strategy of guideline development depends on the acceptance of the recommendations by providers, especially physicians. Physician acceptance of developed guidelines will be facilitated to the degree that there is agreement with the underlying objectives of the program and the actual guideline development process itself. The process must be open, clearly defined, allow for direct clinical input, and be perceived as objectively assessing the information. To the degree that the objective of guidelines is to define optimal clinical practices, guidelines should be developed by a group that is independent of those responsible for insuring or paying for the services in question and that is dominated by a medical-scientific orientation. Participation of professional provider groups will increase provider and patient confidence in the process.

The development of guidelines requires input from physicians with substantial expertise in the clinical practice being assessed, physicians with a broad clinical perspective, and experts in the methodology of assessment, including study design and methodology, data analysis, information synthesis, health outcomes, and costs of care. Thus, in addition to the relevant subspecialists, input must be obtained from primary care physicians, clinical epidemiologists, statisticians, decision analysts and scientists, and economists.

Similarly, the issue of reimbursement given the medical-scientific information is largely a social and political decision. While physicians and other providers have important contributions to make in this area, their expertise is more limited. Reimbursement decisions should be made according to the preferences of those who will benefit from and pay for the decision - patients, purchasers of medical insurance, insurers, and political representatives. Guidelines, when appropriately developed, incorporate both the medical scientific data regarding effectiveness and the social and political values placed on expected outcomes relative to cost.

5.1.2.3. Identification of Target Practices

There are far too many practices and far too many clinical conditions to permit development of guidelines for most clinical services. Thus, any guideline program must have a mechanism for prioritizing and selecting the

practices for which guidelines will be developed from among the many candidate practices.

Guideline development should be focused on practices which have high volume, high aggregate financial costs, high health risks, high health benefits, and high levels of disagreement regarding appropriate use (as manifested by variations in utilization or significant differences of opinion among experts).

Guidelines should be developed for new practices and selectively for existing practices. Since the success of a strategy of guideline development depends on provider acceptance of the recommendations, providers must be represented on the panel which selects practices for evaluation. However, since the objectives of guideline development are multifaceted, representation should be broad, also including payors, insurers, and beneficiaries.

5.1.2.4. Development of Guidelines

Guideline development must address a series of strategic issues:

- (1) Distinctions must be made among the objectives of (a) educating and advising physicians and other providers; (b) developing reimbursement guidelines, and; (c) penalizing providers for provision of inappropriate services. The latter objective requires a different set of criteria that would be more appropriately described as utilization review criteria than as guidelines.
- (2) Provision must be made for sufficient flexibility so as to permit physicians to respond to the legitimate individual preferences of their patients, and to the different availability of resources in their environments.
- (3) Decisions must be made regarding whether the process should be technology- and service-based (e.g., the appropriate utilization of the exercise stress test, percutaneous transluminal coronary angioplasty, or coronary artery bypass surgery) as opposed to patient- and problem-based (e.g., the appropriate diagnostic and therapeutic management of patients with suspected coronary artery disease).
- (4) Standards for evidence and decisions must be specified.

Similarly, it is likely and even necessary that there be different levels of approval of clinical practices, based upon the information available and the perspectives of the parties involved. Specific practices might be determined to meet acceptable levels of safety and efficacy, yet also be determined not to be eligible for programmatic coverage due to the high cost involved (e.g., organ transplants), the absence of an ethical and political consensus (e.g., abortion or genetic engineering), or other factors.

Outdated and outmoded clinical practices are uncommon and account for a very small portion of inappropriate service utilization. Most medical

practices offer some benefit to selected patients in specific clinical circumstances. Thus, guidelines will need to be oriented to both the clinical problems (e.g., What constitutes appropriate care of a patient with chronic, stable angina?) and to the practice in question (e.g., What is the appropriate use of coronary artery bypass surgery?).

The actual methods for the development of clinical guidelines are not well established. Decisions in developing guidelines must be based primarily on data. However, there are different classes of data. Primary clinical data from well designed studies are in short supply and rarely sufficient. Methods to evaluate secondary data (e.g., meta analytic, decision analytic, and simulation techniques) exist and are important sources of information. Because of data limitations, expert opinion must play an important role. However, disagreement and conflict are implicit in any task which involves making decisions under conditions of considerable uncertainty. Methods to resolve such conflicts in the most valid and reliable fashion are not well developed.

5.1.2.5. Implementation Issues

Guideline programs require several levels of programmatic flexibility in implementation. When the objective of guidelines is to educate and advise physicians and other providers and to develop reimbursement guidelines, the burden of proof must lie with proponents of the service. Conversely, while there must be justification available for use of a service, the burden of proof must lie with those denying payment when a provider is being penalized for inappropriate use of a service. Guidelines must be sufficiently flexible so as to permit responsiveness to individual patient preferences.

One method of implementing guidelines as a volume containing method is to require physicians to justify their practices. Utilization of diagnostic tests may be decreased if physicians are required to justify the ordering of a test. While the observed decrease in utilization is likely to be greater if requests which do not meet accepted criteria are rejected, substantial decreases may be observed, at least in the short term, even if justifications are never reviewed and requests are never rejected. However, the long term impact of such interventions is unknown. Most likely, such a program loses its impact over time if it is not enforced by refusal of inappropriate requests. The administrative costs of running such a program may exceed the savings. Moreover, such a requirement is likely to quickly become onerous to providers when applied to the huge volume of physician practices. However, such programs may be warranted for high cost, high risk, elective, discretionary practices; such as major diagnostic and surgical procedures for which there is evidence or suspicion of significant overutilization.

Publicized guidelines offer the opportunity to inform patients of medical standards and to encourage them to become more active participants in and to assume more responsibility for their own care. An informed consumer can act as a check on busy practitioners, suggesting alternatives, inquiring about alternative management strategies, and raising issues of safety, effectiveness, functional outcome, and costs.

Physicians are becoming increasingly sensitive and responsive to patient preferences, both in the performance of tests and procedures and in the selection of therapies. Patients, in turn, have increased their involvement in their own care. Publicity about breast cancer and mammography have led to increased patient demand for and receipt of appropriate, cost-effective screening tests. Patients have been the primary source of demand for less radical, less disfiguring, and less costly treatments for breast cancer. If patients were to become more cost sensitive as a result of higher deductibles and coinsurance, they would be likely to inquire more about costs of and alternatives to the procedures offered. An informed consumer can also assist the physician in opting not to provide a service which offers little potential benefit relative to its costs: a decision the physician may be reluctant to make on his or her own.

5.1.2.6. Effects on Volume and Expenditures

Guidelines in and of themselves probably have relatively little impact on changing physician behavior. A number of programs have attempted to change physician practices through educational efforts, often using guidelines. The results, in general, have been disappointing. National consensus expert guidelines regarding the use of cancer screening procedures (American Cancer Society), immunization (Centers for Disease Control Advisory Committee on Immunization), and selected tests and procedures (American College of Physicians Clinical Efficacy Assessment Project) have been reasonably well accepted by clinicians. However, their success in being implemented clinically has been variable. Those guidelines which have been accompanied by a great deal of publicity and repetition have attained modest degrees of compliance (e.g., ACS and ACIP guidelines).-- Simply making the information available to physicians has been less successful.

There appears to be some asymmetry in the impact of guidelines -- recommendations to increase practices may be more readily accepted than those which suggest decreased volumes of practices. Thus, compliance with recommendations to obtain Pap smears, to test stool for occult blood, and to immunize patients against influenza have been better accepted than recommendations to reduce or eliminate screening chest x-rays and electrocardiograms. However, it is important to note that all guidelines appear to influence a substantial portion of physicians, especially those physicians who act as opinion leaders within their communities. Also, physicians do not uniformly accept recommendations to provide more services, particularly when the procedure involves some risk, discomfort, or substantial expense to the patient. Thus, screening sigmoidoscopy for the early detection of colon cancer and screening mammography for the early detection of breast cancer have been less well accepted than other, simpler, safer, less expensive cancer screening tests and procedures, even when the physician derives substantial financial gain from the performance of the procedure.

Guidelines that have been developed with substantial involvement of the affected physicians and which are well accepted by physicians appear to have greater impact than controversial guidelines developed without significant input from physicians. Significant involvement of professional societies and

their representatives in the development of such guidelines is useful and desirable for expediting acceptance among providers. Neither practicing gynecologists nor the American College of Obstetrics and Gynecology were significantly involved in the formulation of the American Cancer Society guidelines on the recommended frequency of Pap smears. Consequently, there was a general lack of acceptance of the guidelines and, ultimately, they had to be revised.

The most successful uses of guidelines to alter physician practices have combined guideline development with other interventions. The most widely studied and most successful interventions have used guidelines in conjunction with extensive peer involvement and peer pressure, in conjunction with financial incentives, or in conjunction with a combination of both. A number of experiments involving feedback of information to groups of physicians on their own levels of performances, often in conjunction with expert guidelines or in conjunction with development of local guidelines, have resulted in substantial declines in the use of presumably overutilized services. Prostatectomy rates declined in Maine after urologists received feedback comparing their utilization with those of their peers. This feedback was accompanied by outcome data on the procedure. However, this program was conducted by the Maine Medical Association, giving it substantial credibility. Reductions in hysterectomy rates were observed in Manitoba using a similar program. Broad decreases in the use of ancillary services were observed by a Pennsylvania Blue Shield program which compared providers' utilization with those of their peers as part of Pennsylvania Blue Shield's utilization review efforts (Schwartz, et al., 1988).

Given the high level of service utilization which is not supported by the medical literature and expert opinion, an effective program of guideline development, if coupled with other effective interventions, can be expected to result in substantial reductions in the volume of services provided and their associated costs. Based on the literature, it is not possible to provide an accurate estimate of the level of likely savings. However, based on our discussion with those who have been involved in guideline development and use, a savings of 10% of the medical care dollar in direct and induced costs might reasonably be expected without any net loss of benefit (and, even, with some gain in benefit). The savings engendered from the development and application of guidelines are substantial and recurring. The application of guidelines to new practices also offers the opportunity to reduce the rate of increase of volume and costs.

5.1.2.7. Appropriateness and Fairness of Volume Limitations

One of the major attractions of the use of guidelines is that the volume limitations which result will, by definition, be appropriate. Guidelines seek to facilitate the primary objective of the medical care system -- improvement of patient health and outcome in an appropriate, efficient manner. If the guidelines are well constructed, they also will be fair to recipients. However, it is important to emphasize that there are multiple opportunities for bias and inequity in poorly derived guidelines, depending on which

benefits and costs are considered and how they are weighted. Thus, the methodology of guideline development is an important factor.

5.1.2.8. Feasibility of Guideline Development

There is widespread experience with guideline development by a variety of public and private sector groups. There are many examples of guideline development by government agencies (CDC ACIP immunization recommendations) and panels (U.S. Preventive Services Task Force), professional societies (American College of Physicians Clinical Efficacy Assessment Project, AMA Council on Scientific Affairs), insurers (HCFA heart transplant guidelines, Blue Cross and Blue Shield Associations of America Medical Necessity Project, NCHSR-HCTA review for HCFA), private research groups (Rand studies on selected medical and surgical procedures), and non-profit societies (American Cancer Society cancer screening recommendations). The use of carefully derived, objective expert guidelines is widely accepted by providers, patients, and insurers. From the provider's perspective, guidelines help in addressing a complex and vexing issue. They provide expert consensus where none presently exists and may reduce defensive medical practices and malpractice premiums by providing national standards of care. Patients support guidelines because they ensure that cost containment will not come without regard to quality of care. Payors and insurers accept guidelines because they provide a broad-based consensus approach to volume control.

However, the feasibility of guidelines is dependent on the availability of the required data. The limiting step in the development of comprehensive guidelines is an absence of data on the efficacy and effectiveness of clinical practices, particularly as they relate to impact on patient function and outcome and to patient preferences and utilities. There is very little research funded research to collect such information and the practical and economic issues involved make it infeasible for individual providers to collect such data. The development of a nationally based guideline program must be accompanied by a concurrent applied research program to collect the required clinical and economic information and to improve the quality of guideline development methods. Efforts on the scale of a national institute, or at a minimum, a very significant expansion of NCHSR activities, would be appropriate. Other efforts to develop data bases, including the use of claims data for the development of guidelines, should be encouraged.

Developing the data needed to formulate clinical guidelines will be facilitated by stricter implementation of the principle requiring supporting information on the safety, efficacy, effectiveness, cost, and benefit of all new procedures, technologies, practices, and services before they are approved for reimbursement.

There are few medical practices which are totally without value. Rather, the utilization issues surrounding almost all medical practice concern their marginal benefits and their costs. While a great deal has been learned about the identification and measurement of medical outcomes, costs, and the relationship between them, analysis of the cost-effectiveness of medical practices remains controversial. Much more research concerning the

methodology to measure costs and outcomes is required in order to facilitate more appropriate, more efficient medical care and to develop high quality, broadly accepted guidelines that will improve patient health.

Guidelines require tradeoffs between the marginal costs and benefits of medical practices. Research is required to identify, define, and develop methods to measure both relevant health outcomes, and costs. Particular attention must be directed toward patient preferences and utilities for alternative costs and health outcomes and toward the improvement of methods to determine how to tradeoff costs versus improved health outcomes.

Development of guidelines for newly emerging practices offers special opportunities. While the greatest immediate impact of guidelines will be on already existent, high volume practices, their application to emerging technologies and practices is easiest to implement and, over the long term, offers the opportunity for significant savings. There is less opposition to guidelines before a procedure is well established and before it has many adherents and proponents. Moreover, the opportunity exists to set explicit, clear standards in advance. The application of guidelines to heart transplantation has helped control utilization (although the supply of organs may be the rate limiting step) in a much more accepted fashion and with better results than attempts to control the use of coronary artery bypass surgery. Over time, the ethic of adherence to guidelines becomes established and reinforced, spilling over to other practices.

5.1.2.9. Other Benefits

In addition to the direct volume and cost control benefits derived from the development of guidelines, a number of important external benefits accrue. Guidelines may contribute to the reduction of defensive medicine by providing the physician with support for a decision not to proceed with a practice which is unlikely to be of any value. Perhaps most importantly, carefully developed expert guidelines can help build confidence that other components of cost containment efforts will not come at the expense of patient care quality. Thus, guidelines can serve to strengthen confidence and broaden support for concurrent cost containment efforts.

Guidelines also offer the potential in the intermediate and long term to affect provider decision making. A reorientation of thinking toward the consideration of the issues of costs and cost-effectiveness tradeoffs in clinical decision making should result in more appropriate, cost-effective care in the long term.

5.1.2.10. Costs of Guidelines Development

The development of well-accepted, well-crafted, high quality guidelines is expensive. Based on conversations with foundations and organizations that have or are in the process of developing guidelines, the process of guideline development is estimated to be somewhat in excess of \$100,000 per set of guidelines. While many attempts to formulate guidelines are less expensive,

these latter programs rely on substantial voluntary effort and motivation. An institutionalized, mandatory, sustained program will not be able to rely on such voluntary effort. Moreover, the increased importance of governmental standards will lead to increased scrutiny and more complex procedures, further increasing costs. However, relative to potential cost savings, such costs are small for many practices and procedures.

In addition to the financial costs of development, the use of guidelines is associated with a variety of non-financial costs. These include the risks of prematurely entrenching practice patterns into a standard format and thus prematurely freezing the opportunity for future practice pattern improvements; reducing or eliminating appropriate variations in practices in response to patient specific factors and preferences for various alternative outcomes and risks; and adding additional complexity (and antagonism) to what will be more frequent appeals for exceptions to guidelines.

5.1.3. Summary

Scientifically derived guidelines offer the potential to improve provider information and to improve the quality and cost-effectiveness of medical care. In fact, such programs probably are necessary, at some level, to achieve such changes in practice. However, while necessary, guidelines almost certainly are not sufficient in and of themselves to cause substantial reductions in inappropriate utilization in the short or intermediate term. There is ample evidence that information and education alone rarely are sufficient to produce rapid, significant, and persistent change among physicians and other providers. Rather, the most effective interventions depend upon multiple components of education, financial incentives, peer values, and administrative changes. Thus, the development of guidelines should be perceived as an important component of volume and cost control mechanisms, to be integrated with programs to provide financial incentives for cost-effective care (e.g., denial of payment, sharing in savings, and market pressures from patients who share in the cost of care), to reorient providers' perspective (e.g., through professional societies, peer practice comparisons, and public opinion), to include a service's cost-effectiveness in decision making, and to perform expanded utilization review.

Appendix: Theoretical Effects of
Guidelines Under Behavioral Models

Appendix: Theoretical Effects of Guidelines Under Behavioral Models

Clinical Guidelines Under Profit Maximization

Here we assume that clinical guidelines are issued on the basis of empirical research of acceptable quality or on the basis of professional consensus of an authoritative body and that the information contained in the guidelines was not previously known to the majority of physicians. From a professional perspective, then, the physician might be persuaded to adopt the guidelines. The guidelines are disseminated and compliance is audited. The audit results are communicated to the physician. No penalties are imposed for noncompliance.

Guidelines of this sort will be issued for a selected number of procedures, such as high cost or high volume procedures. Two types of guidelines can be distinguished. Type 1: Guidelines recommending substitution of services (x) by another service (y) which is at least as effective as x and paid at a lower rate. Type 2: Guidelines which restrict the use of (x) to a smaller set of patients.

1. Guidelines to Substitute Procedures

A profit maximizing physician will provide the service recommended by the guidelines if its provision results in higher net profits. This could happen if the reduction in the cost of inputs for the recommended service is larger than the decrease in payment.

In the case where the recommended procedure yields lower profits than the procedure currently used, the physician will not have an incentive to adopt the guidelines. However, if physician noncompliance may result in decreased demand in the long run, long term considerations will create incentives to adopt the guidelines. Long term demand for the physicians' services may also be adversely affected if patients knew that the recommended procedure is at least as effective as the procedure currently offered, if the beneficiaries' out-of-pocket costs for the recommended procedure were lower than the out-of-pocket cost for the current service and if there are alternative physicians willing to perform the substitute service.

Consequently, even guidelines that reduce current physician profitability may be adopted. Such guidelines should help decrease the volume of services reimbursed at high prices, but their impact is likely to be observed in the long run. The degree of compliance with the guidelines would depend on the ability to demonstrate to the patients that the recommended procedures are effective and reduce users' costs. The volume of the currently provided procedure will decrease because of the substitution to the recommended procedure. The volume of the recommended procedure will be lower than the volume of the current procedure because the incentive to induce demand for services with lower net profit is weaker, since the value of leisure is relatively higher. Total expenditure should, therefore decline. Quality of care will improve.

Since guidelines may reduce the incentives to induce demand, it would be useful to issue guidelines for overutilized procedures known to have a negative or damaging effect on quality of care when overutilized.

2. Guidelines Restricting Use of a Procedure

In this case, the profit from provision of a service does not change because of the guidelines. The guidelines' effect is to decrease demand and raise the psychic costs of demand creation. That is, the maximum demand patients will be willing to accept will fall. The guidelines make it clear who should not be subjected to a given procedure. The guidelines cause the demand function to move down and the supply function to shift up.

In the long run, guidelines will result in volume reduction of the particular service and decline in Medicare's expenditures. Access to care may be adversely affected if the rise in the supply curve is steep and the marginal cost equals the reimbursement rate before true demand is satisfied.

If guidelines result in excess demand, an increase in balance payment for the service could be observed.

This analysis depends on the assumption that patients know, or are believed to know, the guidelines and the price of the various services.

Under this model, quality of care will improve because of decline in overutilization. Also, since the demand for other services has been fully exploited, no other changes in the system will occur. Thus, total Medicare expenditures will be reduced, volume will decline and quality improve.

General Comments

In general, audits are used to increase compliance with guidelines through professional peer pressure. In the profit maximizing model, professional pressure that has no financial consequence is not admitted as an incentive which will alter behavior. If policy makers would like guidelines to be effective in the short run, and not so dependent on future demand, they can use penalties for non compliance. Penalties reduce profit and will reduce the volume of procedures provided.

Penalties may, however, create an incentive to use balanced billing. Assume consumers are not aware of the quality differential between the recommended procedure and current procedure. Physicians may induce demand by arguing that the current procedure is suitable, but they can not provide it because of the penalty. Some patients may be persuaded to pay the reimbursement differential and the penalty. Thus, patient access to information about price and outcome is crucial to effectiveness of guidelines, as a volume control mechanisms.

Guidelines Under Income Targeting

In the income targeting model there is a stronger inclination by the physician to accept and follow guidelines. By assumption, the physician values the provision of updated clinically reliable information to his patients. We can, therefore, expect a reduction in volume for the procedure found to be inferior or for the procedure whose use has been restricted.

The impact on total volume and expenditure is undetermined. Assuming that compliance with the guidelines results in loss of income, the physician in this model can induce demand for other services or cut the costs of input, thereby reducing quality of care, to increase net income. The decision will depend on the balance struck by the physician between the income effect, the substitution effect and his subjective valuation regarding his professional conduct. Guidelines for very profitable procedures which constitute a large portion of a physician's practice would cause a larger adjustment in the physician provision of other services.

Guidelines will have a positive impact on the quality of care for the specific service or condition to which they are directed. However, the quality of care to other patients or services can decline because of induced demand or cost saving measures undertaken to increase net income.

Clinical Guidelines Under Patient Agency

There is an important conceptual difference between standard-setting for education and optimal clinical care (in which marginal services--those that have a positive but small net benefit for at least some patients--are generally discouraged) and standard-setting for denial of payment or penalty (in which marginal services are generally tolerated and only grossly unnecessary services are disallowed). Exactly how standard-setters should decide how much positive benefit is enough to justify a more costly service is, at present, not known; they do emerge with recommendations but the validity of the recommendations is suspect. The difference is in whether the burden of evidence is on the proponent of a service to prove the service to be better than the less expensive alternative (a problem of alpha error) or on the opponent to prove the service to be no different from the less expensive alternative (a problem of beta error). Here we consider guidelines of the former type, in which appropriate and/or optimal strategies are specified.

If physicians are unaware of the appropriate services to provide, clinical guidelines could aid them in providing services of optimal clinical utility. The effect on the volume of a particular service could be in either direction, depending on whether physicians have previously under-estimated the net benefit of that service or over-estimated it. However, it is most likely that doctors initially provide services of small marginal impact, either because they feel that they want to provide every service that may have positive net benefit to the patient, no matter how small, or because they fear malpractice litigation. (Note that most medical ethicists advocate providing all services of positive net utility to the patient, generally without concern for cost. Standard-setters do implicitly but arbitrarily judge value.)

Therefore, strongly worded guidelines from authoritative groups that deem the net benefit of certain services to be non-existent, or which assure the physician that there is no chance of positive benefit, are likely to lead to less intensive utilization overall than is presently the case.

A number of predictions can be made about the types of guidelines most likely to be successful with doctors who behave as patient agents. In general, the most influential guidelines would be ones that are clearly defined, professionally derived, emanating from respected or prestigious sources, and data based, demonstrating not only that cost can be reduced but that quality can be preserved or even enhanced. That is, they are standards which do not involve valuing service's benefit.

Under the "pure agency" model the more convincing the clinical evidence the more likely is the recommended procedure to be utilized. If the recommended procedure is less costly than previous treatment, expenditure for the particular condition will decrease. Keeping everything else constant, total expenditure will also decrease. However, convincing evidence on effectiveness could increase volumes of the recommended procedure. This is particularly likely to happen if the guidelines identify patients currently not receiving services who should be strongly encouraged to get it as well as identifying some patients for whom the services is currently inappropriate. The impact on resource use and expenditures will depend on the relative savings from reduction in use of the old procedure and the increase in volume of the recommended procedure.

However, if guidelines are not based on clear scientific criteria and outcomes are doubtful, the "pure agency" relationship will predict that physicians will resist acceptance of guidelines. Or suppose the guidelines proscribe a service of undoubtedly positive but small health benefit, which is nevertheless costly to the insurer. The physician-as-agent will resist such attempts to economize at the cost of the health and well-being of his or her population of patients. Policymakers may attempt to counter this resistance by imposing penalties, denying claim payment, and imposing transaction costs on the appeal process. The efficacy of these policy instruments will depend on the accuracy and specificity of the guidelines. The less clear and specific the guidelines and the less definitive the clinical justification for their imposition, the less likely it is that physicians will adopt the guidelines and that expenditures will decrease. If noncompliance rates are high, and transaction costs due to the appeal process are high, total savings in expenditure may be minimal.

Thus guidelines should be selectively used, when they are based on good reliable clinical evidence that can be used to convince physicians to alter their practice.

5.2. Utilization Review

5.2.1. Overview of the Literature

The literature on prospective utilization review (PUR) seems to support the expectation that, if properly implemented, techniques such as prior authorization (equivalent to pre-admission certification in the inpatient environment), second opinion surgery, and case management could successfully reduce volume expenditures.

In a 1974 study, McCarthy et al., evaluated the impact of second opinion surgery for inpatient and outpatient services. The evaluation included two patient groups. The first group consisted of 20,000 self-insured union members for which a second opinion program was mandatory, but not strictly enforced. 602 patients of this group submitted to second opinion and 17.6% were not confirmed. The second group had 200,000 patients, but second opinion was voluntary. 754 patients records were available with a non-confirmation rate of 30.4. The authors estimated that the program's overall benefit to cost ratio was 8 to 1. Ruchlin (1982) also estimated a positive benefit to costs ratio for a mandatory second opinion program applied to inpatient and outpatient surgeries. His estimate, however, is lower at 2.63. Martin (1982) estimated saving of 3 to 4 dollars per dollar invested for a mandatory second opinion program in Massachusetts.

Imperiale (1988) et al., evaluated the impact of a pre-admission certification program for Medicare patients in Connecticut. Out of 28,450 Medicare admissions only 0.37% (n = 100) requests for admissions were disapproved for reimbursement. Some short term minor problems were attributed to delayed admissions, but no permanent adverse effects have been detected. The authors recommend that the impact of delayed admission on medical and psychological patient condition should be evaluated.

Feldstein (1988) evaluated the impact of PU program implemented by a private insurance carrier that employed at least one of the following techniques: pre-admission certification, on-site review and concurrent review. Participation in the program was mandatory and penalties were imposed on patients. An impact was detected resulting in reduction of hospital admissions by 12.3% patients, 11.9% reduction in hospital expenditures and 8.3% reduction in total health costs.

Milstein et al., (1987) reported the impact of pre-certification procedures varies across programs and identified 3 factors associated with levels of program effectiveness. The factors were (1) supportive benefit plans which contain incentives to participate in the pre-certification program; (2) quality of implementation and; (3) lack of objective scientific validation of standards.

In none of these cases was there evidence that PUR reduced the rate of growth in cost beyond its initial impact on the level of costs. This question was addressed explicitly in the Feldstein study. It found that the reduction of total health care costs was "one time"; after the program effect was experienced cost then increased at the same rate in insurance plans with and

without PUR. Thus, there is no evidence that PUR produces a permanent reduction in the rate of growth of cost.

Could retrospective UR (RUR) effectively control the rate of growth in Part B volume and expenditures? While the effectiveness of RUR in controlling inpatient utilization has been studied extensively, the effectiveness of RUR in controlling ambulatory services has not been extensively reported. There is, however, a limited body of literature which evaluated the effectiveness of RUR simultaneously in the inpatient and outpatient settings. This literature seems to confirm the presumption that RUR could be effective in controlling volume of outpatient services.

Buck (1974) evaluated record of 259 physicians providing services to Medi-Cal patients. He found a statistically significant reduction in services. The reduction was not uniform across services. But physicians with higher utilization rates were more responsive to program sanctions and reduced utilization by a higher rate than the physicians which were classified as "normal utilizers. Brook et al., (1978) evaluated Medicaid claims data in New Mexico and found RUR to be effective in reducing volume and improve quality of ambulatory services. Here again, the impact of UR was not uniform across services; office visits were not impacted by RUR but the volume of injections was reduced. Inpatient services, on the other hand, were not affected. Retrospective UR for ambulatory services was also evaluated by Paris et al., (1980) for 7,582 physician providing services to New York Medicaid patients. This study isolated 402 high utilizers. Rate of utilization for 55 physicians who were subjected to program sanctions decreased. Thus there is a set of literature which confirms the proposition that retrospective UR of physician utilization of outpatient services can be effective in producing at least a one time reduction in utilization of some ambulatory services.

5.2.2. Likely Effects of Implementation

Utilization review is a set of techniques used to administratively monitor and control the volume of medical services. Utilization review procedures are based on the comparison of proposed or delivered medical services against predetermined criteria. These criteria can be derived in a variety of ways. They can be based on average utilization patterns within a geographical area or institution or they may be based on normative criteria and at times they may be adopted from guidelines or studies which recommend proper utilization. UR's conceptual justification is rooted in the belief that overutilization accounts for a substantial part of health care expenditures. UR, by establishing practice standards could, it is argued, reduce physician practice variation and thereby reduce total volume and intensity. It would eliminate opportunistic behavior since deviating physicians could be identified, and it could be used in conjunction with an enforcement system as a cost saving mechanism.

It is possible to distinguish among two types of UR procedures, retrospective UR (RUR) and prospective UR (PUR), also known as utilization management. (Another kind of utilization review is concurrent review, however, because concurrent review separate and distinct from prospective

review is rarely used for ambulatory care, it is not addressed separately here. Prospective UR techniques include a utilization review component but stress, in addition, a prospective management component that RUR procedures do not include. PUR consists of techniques such as prior authorization, (equivalent to pre-admission certification in the inpatient setting) second opinion procedures and case management. HCFA has developed retrospective UR procedures and prospective UR procedures for Part A. RUR procedures have been used for a limited number of services in Part B. No PUR or concurrent techniques have been applied by HCFA to Part B claims.

The question considered here is whether RUR techniques and PUR techniques applied to Part B can be effective in controlling volume and expenditures and whether these control mechanisms should be adopted. If they are to be adopted, how should they be implemented to provide maximum return for the cost of operating the UR system?

The likelihood that an effective UR system could be created depends on the ability to overcome methodological, political and administrative/budgetary impediments.

In what follows we: (1) assess the likelihood that the impediments to an effective UR program can be removed; (2) assess the empirical likelihood that the implementable UR program could affect volume/intensity; and (3) consider whether such a program should be adopted and propose an implementation strategy.

5.2.2.1. Prospective Utilization Review

Methodological Impediments

The experience of third party payors and our analysis (see section 5.2.2) indicate that successful implementation of PUR crucially depends on the clarity of the criteria according to which UR standards are set and the ability to demonstrate that PUR standards are at least as good in terms of quality of care as other treatment practices. Methodologically, the problems are not unique to HCFA and center around the choice of an appropriate unit of observation, definition of appropriate and clear criteria and relating the UR standards to quality of care. It is unlikely that these methodological issues will be resolved in the near future for most of the services provided under Part B. A more likely scenario envisions the possibility that proper criteria could be developed for some services and that a long term program can be instituted to study the relationships between various UR standards and their impact on quality of care.

Political Impediments

Unique to HCFA as a governmental agency and as a sole provider of insurance to a special population are some possible political impediments. HCFA must secure the participation of a sufficient number of physicians in the provision of care to its beneficiaries. HCFA must also assure its

beneficiaries that the physicians are provided the opportunity to deliver accessible, appropriate and effective care. These two constraints, coupled with the potential political power of physicians, impose a lower bound on the strictness with which a PUR program can be implemented. For example, what rule should HCFA use to deny payment following PUR? Should HCFA show that a proposed service is detrimental or is it sufficient to show that the service proposed by the physician is no better than the PUR standard? Both patients and physicians must accept the denial criteria if the UR program is to be effective.

Since the effectiveness of PUR standards is likely to increase with the patients' ability to effectively monitor their physicians, it is preferable that information about standards of care and physician performance be disseminated to the public. This component of the system is not commonly used by private carriers. Implementation of this component in an effective way raises difficult problems. The standards must be clear to prevent incorrect application of the information by the patients. Physician performance data, if distributed, is bound to be a difficult and controversial component of a consumer education program. When released, the information could prompt physicians to exit the system in order to protect their non-Medicare market. Release of UR standards to public use may affect the provision of care.

It is unlikely that physicians will mount a fight against a UR effort over Part B. PUR is not a new strategy; it has been employed by private carriers and has been already established as an accepted tool for Part A. Resistance would diminish if physicians were invited to participate in the development of additional PUR initiatives, especially if PUR development would be linked to quality of care studies. However, the strictness with which HCFA could impose penalties and the manner and extent to which physicians' performance data would be disseminated could be contested. UR implementation by HCFA may therefore be less efficient than in the private market.

Administrative/Financial Impediments

Some administrative/financial impediments are also unique to HCFA. Verifiable standards for Part B services which are based on quality of care criteria will require access to Part A data, for some services such as rehabilitation after inpatient care. Access to medical records for outpatient care, as distinct from claims data, would also be required. Assembling the needed information, processing it and verifying that UR standards ensure an acceptable quality of care will be a costly process. Whether HCFA will be willing and able to allocate the funds to undertake such a project is questionable.

The potential benefit/cost ratio of PUR implemented by HCFA is particularly difficult to assess because comparable implementation conditions are not available and the available literature is scarce. For instance, private insurers deny benefits but cannot prevent the physician from billing the patient, as Medicare would typically be expected to do. Moreover, there is relatively little evidence and expertise on which to base either clinical or management standards for ambulatory care (Palmer, 1988). If we assume that

UR for Part B is going to be as effective as for Part A, the dollar savings is going to be lower, since the costs of development and implementation will probably be higher. Higher costs are expected because ambulatory services do not have a uniform coding system. Development and implementation of a comprehensive uniform data set for ambulatory services is likely to be a costly process to develop and implement. Thus the benefit/cost ratio for Part B will be lower than for Part A. If we add to this calculus the observation that only a portion of Part B services could be subjected to effective UR standards and that the possible actual savings will occur in the longer run when reliable and acceptable standards are developed, the willingness to pay today for those future uncertain benefits may be limited.

There is, however, significant interest in developing effective mechanisms for measuring and monitoring the quality of care. PUR packaged as part of the quality of care agenda will encounter less resistance, and could be supported politically by the medical profession, particularly if the medical profession is invited to participate in the development of quality measures.

Although severe financial limitations for the development of a full-blown PUR program of the sort needed to be truly effective are likely, it would seem that the current environment is as amenable as it ever has been to a limited strategic program that is coordinated with an initiative to develop quality of care standards.

We conclude, therefore, that in the near future we cannot expect the implementation of a full-blown and effective PUR program for Part B services. We also conclude that the implementation of a PUR program for Medicare may not result in the level of effectiveness observed in the private market, nor can all the techniques used in the private market be utilized by HCFA.

A Feasible Prospective UR Program

One could envision, however, the development of a long term plan to utilize PUR in conjunction with the development of a quality control system. Such a program will be targeted initially, depending on the available budget toward a specific set of problems. The initial targets should not be selected according to their likelihood to produce significant reduction in volume or expenditures, but according to their potential to be effectively implemented. That is, PUR should be targeted to the cases for which clear reliable standards that are medically justifiable can be formulated, enforced and verified. The strategy should be to establish, in the short run, the acceptability and reliability of the program. This strategy will build the basis for future expansion and effective utilization of UR.

Implementation of such a selected sequential UR program requires that some choices be made regarding the type of review to be developed, the unit of observation for which the standards are to be developed and the type of criteria to be employed.

Developing prospective review procedures would seem a worthwhile strategy because it will allow development of UR standards around a diagnosis or specific medical need. The effectiveness of the standards could be evaluated and in the long run used to modify and refine the standards. Prospective UR has also the benefits of being potentially able to restrict the population to which some procedures are provided and propose substitution of high cost, ineffective services with more effective procedures. Prospective reviews around specific medical needs also provides the potential for development of UR standards around a package of treatments, thereby enabling control of the use of complementary as well as substitute services. Last, and not least, compliance with prospective standards could be tied to financial rewards directed towards the patients (lower coinsurance rates or deductibles), thus directly involving the patient in the decision making process and providing incentive to monitor physicians' decision making.

Physician Response and Consequences of Prospective UR

Assume that prospective review for some specific conditions have been instituted. The effect is to recommend that some procedures should not be used for a portion of the population (restricting perceived demand) and that for the rest of the population a substitute package of services should be provided. We also assume for purpose of the analysis that the standards are accepted by patients and physicians who are willing to comply with them.

First, it is important to ascertain if the physician could offer the recommended UR substitute at the prevailing Medicare price to satisfy the demand.

The impact of UR could be to make the provision of the service, and particularly induced demand, more costly. The standards impose higher subjective costs on induced demand, transaction costs are higher and it is harder to induce demand since beneficiaries are knowledgeable. In contrast, if the prevailing Medicare price for the substitute is high enough, there will be excess supply and attempts will be made to induce demand for the substitute service. Some induced demand may actually occur since the patients who were convinced to receive the initial service unnecessarily may still be convinced to obtain the substitute services.

It should be noted that PUR standards in our example may result in reduced income to physicians. This will occur if volume of the original procedure and its complements decline and if profit from the substitute procedure are lower than profits from the original procedure. This scenario is likely since the substitute procedure was presumably available and known prior to PUR but underutilized. The consequences of the PUR standards depends then on the physician's reaction to the loss of income. Some "target income" physicians may want to compensate for their loss by inducing demand for services not subject to PUR. This is particularly true if Medicare prices for the recommended procedure are not adjusted upwards, since no new sources of income are made available. Thus we may observe an increase in induced demand for some services. However, since the profitability of services provided has declined the substitution effect creates incentives to reduce volume. PUR, by

defining clear standards and making them known to both physicians and patients, creates further incentives to reduce efforts to induce demand for both the recommended and substitute services. Still, it is possible that PUR simply causes inappropriate care to "bulge out" somewhere else, without reducing its total volume, when not all services are subject to PUR.

The likely effects of prospective UR can be specified as follows:

1. Reduction in volume and expenditures of originally provided services.
2. Increase in volume and expenditure for services recommended by UR services.
3. Impact on volume and expenditures of other services will depend on the balance between the income and substitution effect. Some induced demand for other services is possible.
4. Quality of care for the services under UR would improve, provided there is no excess demand. However, if demand creation results in overutilization of damaging procedures, quality in other areas could be reduced.
5. Administrative or transaction costs will increase. [If the program is effective, the increase will be smaller than the savings from reduced volume.]
6. Feasibility: A program of the sort we described is politically feasible. The serious political impediment is in showing its short term positive benefit/cost ratio. -

Recommendations and Conclusions for Prospective Utilization Review

A long term carefully designed prospective UR program should be implemented. Such a program should be an integral part of a quality assessment research effort. The purpose of PUR should be expanded to include evaluations of possible underutilization and ensure that recommended procedures are at least as effective as rejected alternatives. To avoid problems of excess demand, PUR standards should be accompanied when necessary with appropriate price adjustments.

The program should be initially targeted to ensure its acceptability, reliability and enforcement. The justification for the program should not be its current immediate effect. It should be a part of the quality assurance effort with expected long term reductions in volume and expenditures. Since PUR is unlikely by itself to control increases in volume and expenditures of all services provided under Part B, PUR should be therefore thought of as a component of a wider strategy for volume intensity control.

In the long run, as PUR standards become more inclusive and accurate, they will also have a stronger negative impact on physician incomes. Consequently, incentives to supplement income will increase. Physicians could

try to induce demand, reduce costs of services provided, upcode or refuse assignment. Monitoring physician reaction would become an increasingly important component of a PUR program. Of particular importance for UR review is the possibility that recommended services will deteriorate as cost-saving strategies are employed.

5.2.2.2. Retrospective Utilization Review

Impediments to Retrospective UR as a Volume Control Mechanism

HCFA's ability to overcome the methodological, political and administrative impediments with respect to RUR of Part B would seem to be less problematic than for the PUR techniques (reviewed above). Consider first the political impediments. In the first place, HCFA has a long term experience with retrospective UR on Part A. In addition, some precedent for retrospective review of Part B has already been established. HCFA has 13 national screens for Part B and local carriers have developed and have been using additional screens. This experience leads to the conclusion that the political impediments to retrospective review are likely to be less severe than for PUR.

The methodological impediments that apply to PUR apply equally to RUR. The major issues are (1) the relationship between RUR standards and quality of care and (2) the availability of uniform data which could serve to develop standards and verify them.

The administrative/budgetary impediments to RUR will differ in degree from PUR budgetary impediments, but not in kind. Thus the analysis of this impediment presented in the PUR applies here.

5.2.3. Summary and Conclusions for Retrospective Utilization Review

RUR by itself is not likely to reduce the rate of growth of physician services. Therefore, it should be applied to Part B only as part of a global strategy aimed at controlling volume and intensity. The implementation of a RUR could also have an indirect impact on utilization, just because physicians will know that their utilization is monitored. RUR is likely to be more effective if its standards would be based on reliable clinical data. RUR data could be used to verify and update utilization standards, and thereby further contribute to proper utilization of physicians services.

Appendix: Theoretical Effects of Utilization
Review Under Behavioral Models

Appendix: Theoretical Effects of Utilization Review Under Behavioral Models

Utilization Review With Enforcement Under Profit Maximization

The purpose of this analysis is to ascertain the reactions of a profit maximizing physician to the implementation of UR procedures. The presumed impact of UR standards is to reduce the volume and intensity of services by restricting the use of the service to a limited population and by proposing substitutes which are reimbursed at lower rates and are less profitable. Two questions emerge: (1) Under which conditions will the physician comply with the standards set by the UR? and (2) How would the profit maximizing physician respond to UR standards which reduce his profits?

The likelihood of successful program implementation and the physician's behavioral response will determine the program's impact.

In principle, a profit maximizing physician will comply with UR standards if noncompliance results in lower real income because of reduced revenues or increased costs.

Revenues may decline because of imposed penalties. Compliance will depend therefore on (1) the magnitude of the penalties and (2) the likelihood that the penalties will be imposed. If denial of payment can be appealed, the transaction costs of the appeal process and the likelihood of successful appeal will affect costs and therefore the propensity to comply with UR standards.

Noncompliance can also cause long term reduction in revenues because of decreased volume of output. A noncomplying physician may be perceived by his patients as providing substandard care. This is particularly likely to occur if most physicians comply with the proposed UR standards and if patients are familiar with the standards, or if noncompliance results in inferior outcomes. UR standards that improve quality of care when made public increase patients' incentives to monitor physicians' decisions. In addition, volume could also decline if out-of-pocket expenses increase because of payment denials that physicians pass on to patients when physicians refuse to accept assignment.

This analysis has the following policy implications:

- (1) UR standards should be unambiguous and well established. This will contribute to effective enforcement.
- (2) UR standards should have demonstrated positive impact on quality and should be made public. That is, patient follow-up studies should demonstrate that UR standards are at least as effective as other procedures.
- (3) UR will be more effective if noncompliance affects the out-of-pocket expenses of beneficiaries.

If physicians accept assignment, the effectiveness of UR will depend on the program's ability to deny claims and impose penalties.

The substitution of procedures with lower reimbursement rates (which we assume are less profitable) and the reduction in amount of services provided results in loss of income. The physician may attempt to obtain additional income by adopting cost-cutting measures, but in this model such measures would have been completely utilized prior to UR implementation. The other possible adjustment is to refuse assignment. In the long run, this strategy, however, will reduce demand for the physician services.

The substitution effect creates incentives to reduce volume. UR standards are likely to be targeted towards the more profitable and more expensive services where physicians have the strongest incentives to induce demand. If UR is effective, the marginal income from work will decline resulting in the reduction of volume.

Consequently, in this model, the following results will obtain: Volume of services will decline, expenditure will decline. Access to some services may be restricted since physicians elect to reduce their volume. Limited access to some services may, however, have a positive impact on quality of care if the services for which volume is reduced (in addition to those reduced because of UR standards) were overutilized and had adverse effects on quality. Access to the services for which UR standards have been developed will be at an appropriate level.

In the case of excess demand, UR is not really needed. There is no incentive to induce demand at Medicare's low price. In this environment, beneficiaries may have to obtain services privately outside of the system. Raising Medicare prices up to the appropriate equilibrium price will improve access and contribute indirectly to improved quality.

Utilization Review Under Target Income

In this model, the physician's behavioral response incorporates the satisfaction from providing good medical care as well as financial rewards. To the extent that this model better approximates reality, compared to the profit maximizing model, it highlights the need to have good UR standards that demonstrably enhance patient care in order to provide incentives for both the patient and the physician to comply with the standards.

We first assume that physicians are ready to accept the UR standards because they are consistent with improved care. In this model, physicians will be willing to comply with standards that improve physician knowledge and benefit the patients.

The income effect, due to loss of income, creates incentives to supplement income. This can be achieved by inducing demand for other services, not subject to strict UR, that the physician had refrained from providing. Consequently, expenditures and volume for services not included in the UR review, will increase. The impact on total expenditure could not be predicted. The impact of the induced demand on quality of care is likely to be negative since physicians could have previously utilized those services but decided not to do so, because they did not judge those services to be

consistent with proper professional standards. Whether the net effect is to reduce or increase average quality is unknown.

The substitution effect, on the other hand, increases the propensity to reduce volume. The volume of services provided will depend on the individual physician's rate of substitution between income and leisure and his subjective valuation of appropriate medical practice. The higher the loss in income the stronger the incentive to increase volume and sacrifice professional standards.

In this model, the physician may also choose to compensate for loss of income by reducing the cost of inputs and thereby the quality of the services he provides. The extent to which this option is exercised depends on the rate of substitution between income and the subjective costs of deviation from professional standards. Effective UR can therefore have adverse effects on the process of delivery of care where savings in inputs are possible. Such savings could occur in the provision of the services for which UR standards have been developed. It is therefore important to continuously monitor the relationship between the recommended standards and patient outcome.

What are the long term effects (time series) of a UR program? In this model and under the current assumptions, it will become harder and harder to induce demand for the services under review, because UR standards imply better care, a fact known to both physicians and patients. The costs associated with demand creation will increase because of increased patient resistance, increased likelihood of penalties as improved UR standards are developed for a larger number of services, and the need to make larger concessions away from appropriate care. Opportunities for cost savings will be utilized and eventually exhausted. The process will result in larger reductions in income, a loss to which physicians will have to respond.

In the Medicare system the possible responses are refusal to accept assignment and upcoding. Refusal to accept assignment will increase beneficiaries' burden and will affect the access to care of the poor beneficiaries. Upcoding would result in increased expenditures. UR procedures could, however, be implemented to monitor for upcoding.

This scenario is unlikely because UR of Part B is not likely to quickly result in a comprehensive set of UR standards verified and supported by quality of care criteria. This analysis is nevertheless useful because it indicates how some physicians may respond to an effective UR program. The options of upcoding or not accepting assignment may be preferable to some physicians over demand creation or cost-cutting, when these last options become unacceptable because of the individual's commitment to professional standards.

A more likely scenario, especially for Part B, can be based on the assumption that UR standards for some services could improve, but for a large number of services standards could not be developed and demand could be induced. Thus, we are likely to observe a one time reduction in expenditure followed by an adjustment process in which expenditures increase as physicians induce demand for new services.

The analysis up to this point was based on the assumption that UR standards are accepted by the physician as superior to his own practice, and that in the long run this fact will be known to the patients. A more likely scenario could be based on the assumption that the UR standards are disputable and their impact on quality cannot be demonstrated unequivocally.

In this case, the physician will resist complying and would attempt to convince the patient that his recommendations, even though more expensive, are the better ones. This strategy could be employed selectively depending on patient's economic status. Poor patients will obtain the care recommended by the UR standards. Patients of high economic status will be convinced to pay the full charge of the service and incur the penalties, if the physician does not accept assignment.

In this case, total Medicare expenditure will be reduced. The incentive to induce demand will not be as strong as long as a physician does not accept assignment. If the physician does accept assignment then the physician's response to UR will depend on the penalties imposed and their likelihood, and on the physician's willingness and ability to induce demand. The financial burden on some beneficiaries will increase. Care will not be equitably provided since access to some services will be denied to the poor. The impact of this limited access on quality, is by assumption, unknown.

Utilization Review Under Patient Agency

If physicians truly are their patients' agents, and they believe that the development of standards or feedback of utilization review are better than their own individual judgement, then they are likely to be responsive to feedback that may help them to determine when their practices deviate from accepted patterns of care. In the case of the agency model, the size of the penalty imposed on the doctor for good standards is irrelevant, since doctors will always comply in any case. In order for the standards used in UR to be convincing to physicians under this model, the standards would have to derive from clinical guidelines developed by professional leaders; statistical norms of practice developed by insurers are not like to be persuasive. Physicians are likely to be responsive to utilization review when it helps the physician convince patients of the most appropriate course of action (e.g. discontinuing futile treatment). However, utilization review can also be viewed by physicians as a constraint on or interference with practice if the UR standards are not accepted as valid. In such a case, this model would predict significant resistance to UR decisions by physicians who are already practicing in a way that they believe is optimal for the patient. As for UR based on clinical guidelines, the impact of utilization review on volume is ambiguous, but most likely would result in an overall decrease if physicians agree that wide variations in utilization rates, as demonstrated by Chassin, et al., indicate significant amounts of inappropriate care.

The most effective utilization review efforts would be ones in which personal feedback is provided to the physician, as is the case with second opinion programs and subsequent review. As discussed above, effects of utilization review would also be strongest if peers and professional leaders

are involved in setting parameters and providing feedback, and if physicians believe this collective judgement to be better than their own in isolation. A sentinel effect may make statistical detection of the actual influence of utilization review on volume difficult.

5.3. Copayments

5.3.1. Overview of the Literature

The best evidence available on the effect of beneficiary copayment on the use of services comes from the Rand Health Insurance Experiment (HIE). A detailed description of this study can be found elsewhere (Newhouse, et al., 1981); briefly, the study was a controlled trial of alternative health insurance policies that ranged from free care to those imposing a 25, 50 or 95 percent coinsurance, up to a maximum dollar amount depending on participating families' incomes. Over 2,500 families and 7,700 persons from six areas of the country participated in the study for three to five years. The results of this study showed that as the extent of coinsurance the beneficiary is responsible for falls, total expenditures on medical care rise (Newhouse, et al, 1981). Averaged across all sites, expenditures per person in the 95 percent coinsurance plan were 69 percent of those in the free-care plan. Put another way, free care resulted in an increase in expenditures of nearly 50 percent.

Data from the HIE allowed for analysis of the effects of coinsurance on various components of care as well. In each site and year analyzed, expenditures on ambulatory services rose as coinsurance fell (Newhouse et al, 1981); variation in quantities of services consumed accounted for most of this difference among plans. The likelihood of hospitalization was lower in the one plan that had a 95% coinsurance for outpatient expenses but provided inpatient care at no expense to the beneficiary than in the free care plan; the authors suggest that such an effect may be because physicians less frequently see people as outpatients who have illnesses that might lead to hospitalization (Newhouse et al., 1981). Overall, adults who received free care were hospitalized 31 percent more often than those in cost sharing plans (Siu et al, 1986).

Adults with no cost sharing used 86 percent more antibiotics in the ambulatory setting than did those on the cost sharing plans, resulting in expenditures that were about 60 percent higher (Foxman et al., 1987). No difference was found by coinsurance in the charge per prescription, suggesting that the effect of cost sharing on antibiotic use comes about primarily through a reduction in visits rather than as a result of reduced prescribing per visit (Foxman et al., 1987). Overall drug use in the ambulatory setting increased with decreasing coinsurance at a rate similar to increases in total per capita expenditures. All cost sharing plans showed significantly lower drug expenditures from the free plan (Liebowitz et al., 1985).

Persons with no cost sharing had emergency department expenses that were 42% higher than those persons on the 95 percent plan and approximately 16 percent higher than those on the plans with lower cost sharing (O'Grady et al., 1985). Nearly all of these expenditures were due to the decision to use any emergency department services, as opposed to using greater amounts of service once emergency care is sought.

Evidence suggests that cost sharing does not always or even usually selectively decrease care labelled as "not medically appropriate" by a

particular utilization review scheme; in a review of the appropriateness of hospitalizations of the HIE population, Siu et al., (1986) found that cost sharing decreased admissions labelled as "appropriate" and "inappropriate" by equal proportions. Increasing emergency room use for less urgent conditions was, however, observed in the free care plan as compared to the 25 percent coinsurance, suggesting that some deterrence of less appropriate care may occur (O'Grady et al., 1985).

5.3.2. Likely Effects of Implementation

There is strong empirical evidence that larger out-of-pocket payments are associated with lower total levels of physician expenditures and volumes of services (Newhouse et al., 1988). Compared to a situation in which the consumer pays nothing out-of-pocket, deductibles and coinsurance can reduce total spending by a substantial amount.

Without a system of "self-responsibility" in which beneficiaries pay some part of the cost of care, patients will seek all care that they believe promises some potential benefit, no matter how slight and no matter what the cost, and physicians will not hesitate to recommend such care, since the patient has no financial reason to refuse. As will all incentive devices to limit volume and intensity, copayments must be used with great care; the more so because they seem to be more effective than any other device, and because they initially impact beneficiaries rather than doctors. This section explores options for making better use of copayments in the Medicare program.

The thought that copayments could help to limit cost and defer services of low value was presumably what motivated Congress to require a 20 percent copayment for most Medicare Part B services. However, despite the presence of this feature, Part B expenditures have grown recently at a rapid rate. The evidence of current (time series) growth does not necessarily mean that copayments (or other cost-control devices) are ineffective; it may well be that they make expenditures lower than they would otherwise be, but that this was and is a "one-time" effect and is offset by other more recent incentives driving cost upwards. Nevertheless, one might consider ways of using copayments as a method for reducing expenditures and volume further. In this section, we explore alternative ways of making better use of copayments.

The most obvious way of using copayments to produce further reductions would simply be one of increasing the level of copayments. Despite the possibility of offsetting increases in the purchase of Medigap coverage, this would probably reduce Part B spending substantially. Taylor et al., (1988) estimate, for example, that an increase in the Part B copayment from 20 to 25 percent could reduce Medicare outlays by as much as 11 percent. Since the decline in outlays is greater than the decline in coverage, raising copayments must result in an decrease in total volume and intensity, as well as Medicare payments.

However, the disadvantage of increasing copayment is obvious: requiring the patient to pay more increases beneficiary exposure to risk, as well as reducing the value of Medicare benefits. For some beneficiaries bearing this

higher cost would be very difficult and they might be deterred from seeking services that are judged to be appropriate. Are there alternative, less objectionable ways of redesigning the Medicare program to make better use of copayments?

5.3.2.1. Making Copayments Into Effective Cost-Control Devices

There is a reason why existing copayments might not be very effective in controlling Part B cost. The fact is that, for the great majority of elderly, the net out-of-pocket payment is zero despite the existence of Part B copayments. The reason is that more than 70 percent of all beneficiaries purchase private Medigap insurance and this insurance almost always covers the full amount of copayments. Combined with Medicaid coverage for low income elderly, the result is that only about 15 percent of Medicaid beneficiaries are actually subject to out-of-pocket payments.

Moreover, the growth of Medigap coverage over the years has meant that the real out-of-pocket unit price for Part B services has not increased very much and has not increased rapidly relative to the average real income of Medicare beneficiaries, even though real medical costs have increased rapidly. Medigap coverage insulates most Medicare beneficiaries from the effect of increases in the cost of medical care, despite the intent of the framers of the Medicare program that copayments should enlist beneficiaries in the fight to keep expenditures down.

To be sure, not all elderly buy private Medigap coverage. Very low-income elderly have their copayments covered by Medicaid; we will not consider changing this situation. Private Medigap coverage, in contrast, is positively related to income or wealth. The well-to-do elderly are almost entirely covered, while those beneficiaries just slightly above the Medicaid eligibility maximum are least likely to have Medigap coverage, and most likely to be subject to positive out-of-pocket payments. This seems to be less than equitable; it stems in part from the all-or-nothing nature of Medicaid, and in part from the fact that insurance coverage in general is a normal good - one whose purchase rises as income rises. The effect of simply increasing copayments, though it would deter Medicare expenditures, would be concentrated on the low-income elderly.

It might be argued that it is the relatively low-income elderly who will be most responsive to out-of-pocket payments in any case, so that the current pattern of coverage -- given that 70 percent plus will have Medigap coverage -- is most conducive to constraining expenditures. However, it is also possible that the low-income elderly, deterred from using Part B services by uninsured copayments, put off care which eventually leads to more serious -- and more expensive -- illness. In the Rand Health Insurance Experiment, increases in copayments led to decreased use of some preventive services. Conversely, it is likely that the services of marginal value which purportedly have fed the growth in Part B expenditures tend to be concentrated more among the high-income elderly with Medigap coverage.

With regard to those people who do buy Medigap coverage, the situation is, in a sense, the worst of all worlds. Copayments do not serve their intended function of controlling volume and expenditures, because they are rendered ineffective by private coverage. But the coverage itself, because it is often sold as individual coverage, rather than group coverage, has very high administrative cost. Administrative costs often account for as much as 50 percent of Medigap premiums. Were it not for the equity consequences, efficiency would be improved and administrative cost lowered by abolishing Medicare copayments for high income elderly, since Medicare could surely provide coverage at lower administrative cost. In short, the current situation is one in which Medicare copayments do not deter expenditure but do add to the cost of paper-shuffling that society has to bear.

The implication for the use of cost sharing as a way of controlling volume and intensity is therefore obvious. Before we consider raising copayments and reducing current Medicare Part B coverage, we should examine ways of reducing the extent to which the incentive effects of copayments are currently blunted by the purchase of Medigap coverage. These ways should be designed to preserve equity, and should not lead to inappropriate deterrence of use.

One peculiarity of current Medicare policy is the complete absence of copayment for laboratory tests. The use of such tests has been growing rapidly, and it is difficult to see a reason for a lower copayment for lab tests than for other types of services.

5.3.2.2. Efficient Medigap Coverage

Putting the problem in this way immediately raises, again, the question of social objectives. If one's only objective was to cut Part B spending, simply prohibiting Medigap coverage (e.g., by denying Medicare benefits) would be an effective way to do so. But it is obvious that beneficiaries who buy Medigap coverage get some benefit from buying that coverage, benefit that would be lost if coverage were forbidden. To determine an appropriate strategy, we should investigate the determinants of Medigap purchase, and ask whether those purchases represent efficient market choices.

Research on this subject strongly suggests that current levels of purchase, or the level of purchase that will follow accommodation to catastrophic coverage, are likely to be excessive. The reason is that there are currently at least two strong implicit public subsidies to Medigap coverage. (These incentives will be described in the section which follows.) Not only do these subsidies presumably lead to the purchase of Medigap coverage which, at the margin, is not worth what it costs, but also the pattern of subsidies is one in which larger benefits are provided, on average, to higher-income people. Thus both equity and efficiency could be served by reducing or eliminating these subsidies. Not only would the drain of resources into administering private and public insurance be cut, but the level of Part B expenditures could fall substantially as well. Moreover, in contrast to the price-based or incentive-based volume controls, this strategy does not suffer from empirical or theoretical ambiguity as to the direction of

outcome. Eliminating subsidies to Medigap will reduce Medigap purchases, and reduced Medigap coverage will reduce Medicare volume and cost. How much this strategy will cut costs is less certain, but there are reasons to believe that the amount could be substantial. If action against rising Part B payments needs to be taken now, doing something about excessive Medigap coverage would be a promising strategy.

We have thus far discussed Medigap coverage of copayments. Only a few Medigap policies cover any balance-billed amounts. The MAAC limits the amount of balance billing, but still allows it to exist. To the extent that reduction in Medigap coverage would lead to less coverage of balance-billed amounts, that would also reduce the level of volume and intensity. Moreover, since high income people are more likely to buy Medigap coverage, discouraging such coverage would actually be a step toward equalizing access to physicians who do not accept assignment.

5.3.2.3. Why Medigap Coverage is Excessive

One of the subsidies to Medigap coverage actually comes from Medicare itself, and represents exactly the fiscal consequences of the incentives we have been discussing. Consider the consequences of an individual's decision to buy Medigap coverage on Medicare's total spending. If the person buys the coverage, the user price of care falls to zero, and total use of medical services rises. Some of this additional use is covered by the Medigap insurer, and is built into the Medigap premium. But about 80 percent of the cost of the additional Part B use is covered by Medicare. There is no mechanism to translate the message of this increased cost back to the potential purchaser of Medigap coverage.

Ideal insurance coverage, in any market, should reflect a tradeoff between the benefits of risk reduction and the total cost of the increased expenditure induced by that coverage. The Medigap purchaser sees only 20 percent of the true cost of the additional use induced by buying Medigap coverage; the other 80 percent is simply absorbed by Medicare. Another way to say the same thing is that the benefits of additional risk reduction, and the value of the additional medical services whose consumption is induced, are both received 100 percent by the Medigap purchaser, who pays only a fraction of their cost. Even if the risk reduction benefits are only of marginal value, and the benefits of additional medical care are minimal, the person may still choose the Medigap coverage.

Only Medicare's insurance operates in this fashion. In conventional private insurance markets, sellers of "basic" coverage (usually in a group setting) will recognize the need to raise premiums for that coverage if supplementary coverage is added. Buyers ordinarily buy all coverage from the same insurer, so this additional premium automatically gets built into the premium for a policy of greater coverage. Only Medicare fails to take account of the purchase of important supplemental coverage by adjusting its own premium. Of course, even in the private sector if the buyer of supplemental coverage can keep that fact a secret from the basic insurer, no premium

adjustment will be made. "Coordination of Benefits" provisions exist in part to prevent this from happening.

It is easy to see how to deal with the problem of the Medicare subsidy. For those who buy Medigap coverage, the Part B premium should be adjusted upward to reflect the higher values of expected claims.

The average increase in premiums that would be paid could be substantial. Taylor et al., (1988) estimate that Medicare benefits are an average 39 percent higher for those who buy a typical Medigap policy than for those who do not. Unless there is evidence that the size of this effect varies with other characteristics of beneficiary or policy, and assuming that this impact on expenditures is consistent across Parts A and B, an efficient program would be one that imposes a surcharge on the Part B premium of 39 percent for people who buy Medigap.

Should a Medigap policy, provided individually or in an employment group, be able to demonstrate effective cost containment, this surcharge might be reduced. But, in general, such a surcharge would be efficiency-improving. And regardless of the effect on Medigap coverage, the additional surcharge revenues would help to offset rising Part B cost.

As noted above, the likelihood of buying Medigap coverage is strongly related to beneficiary income or wealth. It therefore follows that the Medigap subsidy that Medicare pays is also distributed in a "digressive" way, with benefits that actually increase as income increases. On average, equity would also be improved by increasing Part B premiums for those who buy Medigap.

For purposes of this discussion, the most important impact of this corrective measure would be an expected decline in the number of people who buy Medigap coverage. If people decide to forego coverage when faced with a net premium (Medigap insurer premium plus supplemental Part B premium) that reflects the true additional expense caused by that coverage, it must be because, in truth, the benefits from that Medigap coverage are not worth their cost. Consequently, society is made better off if no purchase occurs, since the value of the resources saved is greater than the value of the benefits lost. Since many allege that Medigap purchasers are more likely to overestimate the value of such coverage, imperfect consumer information will not offset this conclusion.

The other subsidy to Medigap coverage is paid to those whose coverage is furnished by their former employer as part of post-retirement health benefits. We refer here to benefits for retirees who are also eligible for Medicare, not to benefits for early retirees. Employers' payments for such group Medigap coverage is a tax deductible expense for the employer, but is not counted as part of taxable income for the retiree. In contrast, elderly people who do not have employer-provided coverage must pay for their coverage out of their income, at least some of which is taxable.

Here again, the tax subsidy is greatest to higher-income retirees, both because their employers are more likely to provide coverage and because the

taxes they avoid are collected at a higher rate. It is true that employer-provided payments for Medigap coverage are not treated as favorably for tax purposes as employer payments for pensions or annuities, since in the latter case employers can shield from taxation the interest accumulated on prefunded amounts. However, in contrast to the treatment of pensions in which the income from the tax-shielded contributions are treated as taxable income when received by the retiree, Medigap benefits are not treated as part of taxable income.

With the imminent publication of new accounting rules concerning post-retirement benefits by the Financial Accounting Standards Board, employer-provided Medigap benefits may be somewhat restricted in any case, although there is as yet little evidence of substantial cutbacks. There may be a basis for preserving the tax subsidy to employer-provided benefits to early retirees, if we wish to preserve neutral incentives for such retirement. But there is no obvious case for subsidizing Medigap benefits, given the presumption that the current level of Medicare benefits is an adequate representation of social concern for the level of insurance coverage elderly people should have.

The solution to this dilemma is to reduce the tax subsidy. This could be done most appropriately by taxing the value of employer-provided Medigap coverage as part of the retiree's income. Such a strategy would make such benefits less attractive for all workers. But the higher the expected income level of a firm's workers, the more elimination of tax subsidy should discourage purchase of inefficient coverage.

5.3.2.4. Moving Toward Effective Copayments

Using copayments as a strategy to control the growth of Part B payments appears on the surface to be less politically attractive than some other options. In contrast to schemes which initially place the burden on providers, this strategy places the first-round benefit reduction on retirees. Given the relatively high average level of income of elderly households, sharing some of the burden with the non-poor elderly is not obviously inequitable--but it is sure to be unpopular. On the other hand, of all the cost-containment devices we know about, the copayment strategy is the one for which we have the most unambiguous empirical evidence on effectiveness. And it is the one with the strongest demonstrated effectiveness--an effect on expenditure which is nearly twice as great, in the Health Insurance Experiment, as enrollment in an HMO. The ability of physicians to create demand is obviously inhibited when their customers must pay more out-of-pocket. Moreover, if the level of coinsurance in Part B really does represent a level which we regard as appropriate for some elderly people, those who choose not to buy or who cannot afford to buy, it is not logically consistent to argue that there is a general social interest in maintaining Medigap coverage.

It clearly would be possible to achieve greater equity by treating subsidies for Medigap coverage differently for different elderly households with different incomes. Elderly people with incomes close to the poverty line

might be exempted from the supplemental Part B premium we have discussed, and they might be permitted to receive employer-provided benefits tax free; either of these subsidies could decline as income rises. But high income elderly -- say, the 25% of the elderly with family incomes in excess of \$25,000 -- might well be appropriately subject to the tax and removal of the tax subsidy. The recent income-related Part B premium for catastrophic coverage might serve as the basis for an income-related surcharge.

If these strategies do reduce the purchase of Medigap insurance, and if that in turn reduces the level of volume and expenditures, control will have been achieved. Even if the response of insurance purchasing or Medicare use turns out to be small, at least additional revenues will have been raised, both for Part B and for federal general revenues. If appropriately tailored to ability to pay, this strategy to regenerate existing Part B copayments cannot fail to do some good.

Still another modification might be to permit Medigap to be untaxed, and group coverage to remain tax-free, if Medigap benefits are of a type which do not reduce the current Part B copayments. Medigap coverage for nursing home care, for example, could be permitted without triggering a Part B surcharge. By reducing the subsidy for inappropriate Medigap coverage but continuing it for what is regarded as more appropriate coverage, even current subsidy recipients among the elderly need not be penalized financially.

5.3.2.5. Objectives and Design Issues

It is obvious that public policy has been of two minds about Part B copayments and Medigap coverage. The argument that copayments help to deter use of marginal value is presumably the basis for including copayments in Part B coverage. And yet, there have been serious policy steps to modify Medigap coverage in ways that diminish the amount of cost-sharing. The Baucus amendment in fact requires that an approved policy wipe out the copayments.

The reason for the policy ambivalence is probably a fear that the use which copayments discourage may sometimes be use which is appropriate, in some sense. In practice, however, any other control device runs the same risk; changing physician prices, limiting use by utilization review, or moving toward capitation can also lead to mistakes. The question, as yet unanswered, is whether copayment would be more likely to discourage appropriate use than would other alternatives of equal effectiveness when applied in actual (rather than idealized) settings.

There are several possible answers to this concern of inappropriate deterrence. The first is that if a Medigap surcharge on Part B premiums was related to income, exposure to copayments need not increase among low income elderly who would probably be most at risk for medical underuse. One might also note that physicians would presumably have an incentive to try to discourage underuse, by offering accurate and persuasive information about the benefits of care. It would also be possible, though administratively more complex, to retain the subsidy for types of services for which underconsumption prevents a serious danger.

There is also an important point about logical consistency in the Medicare program. Absence of Medigap coverage discourages use now, and Medigap coverage purchase is an inefficient and costly way to buy coverage against copayments. It is not logical public policy to assert that copayments discourage appropriate use, and then to permit copayments to remain in the Medicare program. If copayments do that much damage, they should be abolished for everyone, not (in effect) just for those who can afford Medigap coverage.

It would therefore be possible, at some administrative cost, to waive copayments for types of services for which underconsumption is feared. And the prospect of underuse remains only a possibility, with resolution depending on a better definition of appropriate use and research on the consequences of Part B copayments.

5.3.3. Summary

We probably know more about the effectiveness of copayments than about any other device which could control volume and intensity. We know that they work, though we do not know whether they slow down the rate of growth in spending over the longer term. We also know a little about their effects on use, enough to know that the effect is not perfect. Compared to idealized forms of other controls, copayments will not look as good. Compared to the uncertainty about the effectiveness of other devices, and the possibility that they too may discourage provision of appropriate services, revitalizing copayments may be a reasonable option. In the world of uncertain policy, at least reducing the subsidy to Medigap purchases by non-poor elderly could be reconsidered. With serious effort at quality assurance, copayments could be a good way to limit volume and intensity. -

Appendix: Theoretical Effects of
Copayment Under Behavioral Models

Appendix: Theoretical Effects of Copayment Under Behavioral Models

Copayments Under Profit Maximization

How do copayments affect the profit-maximizing physician? The copayment will only affect the quantity of services demanded by the patient. In contrast to the target income and behavioral models, a change in the coinsurance level, ceteris paribus, will not impact the behavior of physicians. The addition of a copayment will raise the price of the service to the patient, resulting in a reduction in the quantity of services demanded. If the copayment is introduced with no change in the payment to the physician method, the quantity the physician desires to supply will not change. Whether the reduction in quantity demanded produces a decrease in the actual number of services provided will depend on whether the price faced by the physician (the Medicare payment) and the price paid by the patient (the out-of-pocket expense) had resulted in excess supply or excess demand before the introduction (or increase) in the copayment. If there was excess supply already, an increased copayment will reduce the volume of services provided in the market (as discussed above, we assume the profit maximizing physician maximizes demand creation both before and after the increase in the copayment). If there had been excess demand, the result is ambiguous. Quantity demanded may decrease but still be in excess of the quantity physicians are willing or able to supply, resulting in no change in volume. If the change in quantity demanded is large enough to more than offset the previous excess demand, a drop in volume (smaller than the decrease in quantity demanded) will occur.

As copayments apply to all services, there is more potential for volume declines for the services which have a higher demand elasticity. Thus a shift in the relative quantities of services may occur. The relative impact on resource utilization then may be more or less than the impact on volume, depending on whether more or less resource - intensive services have larger declines in volume. In either case, resources utilized will decline. Expenditures by Medicare and in total would, ceteris paribus, decline.

Quality of care may be negatively impacted because of undertreatment or improved because inappropriate services are discouraged. It could also improve because overutilization might diminish. The net impact of increasing of copayment will depend on the relative demand elasticities of each type of service. To minimize adverse impact on access and quality, policies that will restore access to the poor should be established.

Provider equity will not be significantly affected by a copayment. It may be that a larger decline in the services with greater demand elasticity will effect some types of providers more than others.

From a feasibility standpoint, significant additional increases in copayments are likely to be unpopular with both the Medicare population and politicians.

Copayment Under Target Income Target Model

In the income targeting model the physician derives subjective benefits from providing accurate information and good care to the patient. Therefore, the opportunities for demand creation and adjustment in cost of provision of services are not exhausted and to the extent that high copayments could affect physicians' income, we can expect physicians to react. In general, an increase in copayments will reduce quantity demanded because beneficiaries' price has increased. In addition, demand creation will require more effort. Thus, the perceived costs will cause the supply function to shift upwards.

Two cases can be envisioned.

Case One: The true demand at initial Medicare price is D_I (see Figure 1). At this price, physicians are ready to supply D_A . No demand inducement occurred. If the shift in demand because of higher copayment is smaller than $D_I - D_A$, no reduction in volume will occur. If the new true demand at higher copayments is smaller than D_A , D_I then induced demand may occur. Let D_T be the new true demand. The costs of supplying services at a volume higher than D_T will be higher and is represented by the supply curve S' . Assuming physicians are willing and able to induce demand, the new volume will be D_I . The new supply function in this analysis incorporates the physician's subjective and objective costs of inducing demand.

This analysis indicates that an increase in copayments may result in demand-creation for at least some services which were initially underutilized. The impact of such a change on quality of care is indeterminate. Access to care at the new user price has improved. However, given the new user price, overutilization occurs which may result in reduced quality of care. The economic analysis makes demand a function of price, and quality of care is defined by the consumer's desire to purchase services at the given price. In order to ascertain the actual impact on quality, comparison of outcomes before and after the increase in copayments will have to be performed.

Case Two: D_T represents the true demand (Figure 2) and D_I the induced demand. The goal of the increase in copayment is to minimize overutilization. At the new copayment level the true demand is D_T and the new induced supply function is S' . The new volume would be D_I . D_I may exceed D_T (the initial true demand) or be smaller than D_T depending on the extent of the shift from D_T to D_I and on the slope of S .

It is worth noting that in both scenarios a possible outcome is the reduction of volume. Reduction in volume implies loss in income since Medicare prices to physicians have not changed. Physicians can respond to this reduction in income by attempting to reduce costs of inputs. Such adjustments are possible in our model if it is postulated that the subjective valuation of professional conduct includes also considerations of quality of service and not only appropriateness or necessity of services. Under this assumption cost modifications will have an impact on quality of care and on the volume.

The new supply function will be lower than S_I but the service it represents will be of a different quality.

To the extent that increase in copayment result in reduced volume, access to care (especially for the poor) will be limited. The danger is that they will not come to seek care. However, if they would come to seek care physicians could direct the reduction in volume to those who least need the services, thereby improving quality of care.

Copayment and Deductibles Under Patient Agency

Here the agency model is complex. The doctor may be the patient's economic agent and be reluctant to advocate a service when the beneficiary payment increases. However, the service may be of some small positive net utility, so the doctor as clinical agent will want to advocate the service. The physician is forced, in this case, to make a cost-benefit determination on the patient's behalf.

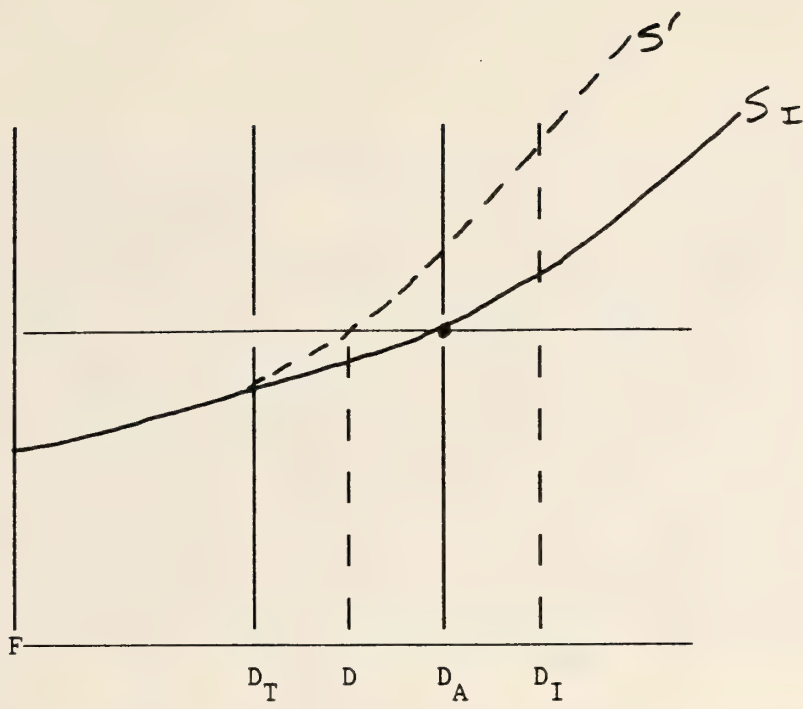


Figure 1

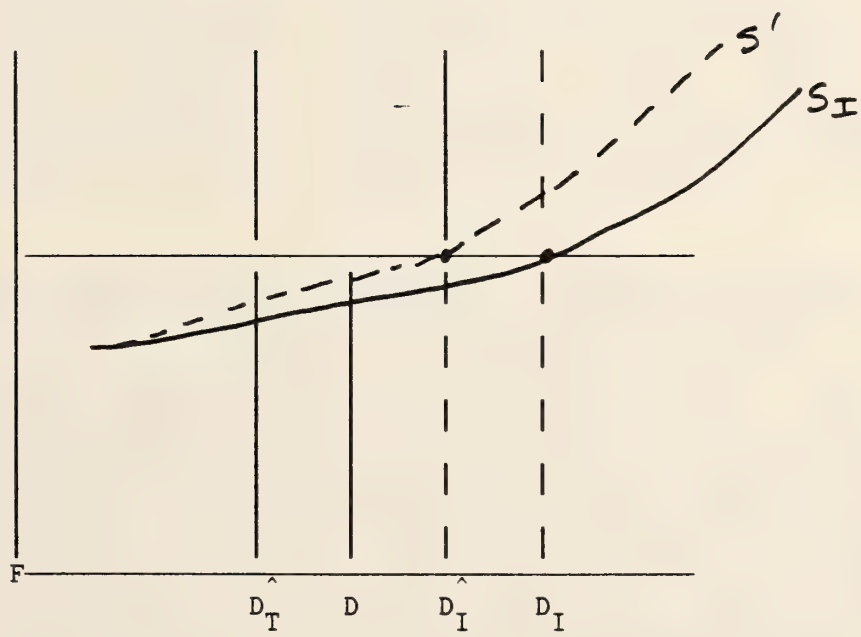


Figure 2

5.4. Capitation

5.4.1. Overview of the Literature

Early evaluations of the effectiveness of capitation in containing volume and expenditures suggested that total costs per capita were anywhere from 10 to 40% lower in capitated plans than in fee-for-service populations (Luft 1980, 1981). However, these evaluations were complicated by the fact that capitated populations tend to be self-selected and thus potentially healthier than those paying for care on a fee-for-service basis. Manning et al. (1984), through the Rand Health Insurance Experiment, reported on a randomized, controlled trial of the effect of a prepaid group practice on use of services. The results of this study indicated that expenditures for the capitated population were 28% lower than for a fee-for-service group that had no cost-sharing. The results comparing the HMO with cost-sharing plans were mixed, with no overall pattern of significant cost advantage for either method of payment. In examining the reasons for this difference, it was found that savings were primarily due to fewer admissions and hospital days overall.

In a comprehensive review of the literature on the relationship between payment method and practice, Hornbrook and Berki (1985) report that several studies have demonstrated this lower utilization of hospitals in capitated plans (Luft, 1981; Sorenson et al., 1981, Blumberg, 1980). They also examined the evidence on costs per service, per contact and per episode for capitated plans versus fee-for-service. They found that costs per patient day or per office visit were not significantly different, and little evidence of differences in the use of technology. They found no studies examining the "typical" bundle of services provided per visit between the two payment methods, and no studies of episodes of illness in this context. They did, however, find that members of HMOs do tend to have more ambulatory visits per year. Thus, they conclude that substitution of ambulatory care for hospital care is occurring in HMOs, and accounts for the overall savings achieved by HMOs.

One of the determinants of the effectiveness of capitation is the financial incentive offered to physicians, which is intended to influence their patient care decisions. The experience of the SAFECO Insurance Company's attempt at capitation through an independent practice association suggests the importance of this financial incentive (Moore, et al., 1983). Using primary-care physicians as gatekeepers (or managers of care), and putting them at limited financial risk, SAFECO hoped to achieve significant cost savings. In a review of why the program failed in this respect, the most salient fact was thought to be the small financial incentives that were used (i.e., limiting the physician's financial risk to 10% of fees on his or her IPA patients, which typically comprised a small proportion of the physician's practice). The plan removed cost constraints on enrollees (as most capitated plans do) without imposing significant constraints on physicians (either financial or in terms of utilization review); the gatekeeper function simply was ineffective in controlling costs in this context.

Some have argued that the effectiveness of capitated programs is more dependent upon a group practice effect rather than the prepayment effect.

Differences in services provided in independent practice associations as opposed to prepaid group practices have suggested such effects (Scitovsky and McCall, 1980; Hornbrook and Berki, 1985). Reasons why this may be so include the evolution of peer standards within a group, and the staffing practices of a group (i.e., a group may not hire additional specialists until the demand is great enough).

Whether capitated plans have been successful in moderating the rate of increase in costs or have simply achieved a one-time savings is also of concern. In 1980, Luft reported his analysis of data from several HMOs and comparison groups over a 25 year period. He found that trends in utilization were comparable between the two groups, so that the rate of growth in total costs was only slightly lower for persons in HMOs. An extension of this analysis for the period 1976-1981 (Newhouse, et al., 1985) confirmed that HMOs cause a one-time reduction in cost, but not a moderation of increases in costs over time.

5.4.2. Likely Effects of Implementation

The conceptual case for capitation as a method of payment that will help to constrain medical costs is strong. By setting the level of payment per person prospectively, it is definitionally true that total expenditures will be affected only by the number of persons at risk, not by variations in the volume and intensity of services provided. The incentives for the recipient or manager of the capitation payment actually to control costs are strong, since every dollar of cost reduces net income by a dollar. And Medicare can save money in three ways:

1. It can set the capitation rate below the expected cost under the alternative system.
2. It can constrain the rate of increase in the capitated amount.
3. If actual cost savings occur for providers, the rate can be reduced to capture some of them.

While Medicare is moving aggressively but carefully toward capitation payment for the full bundle of covered services for those beneficiaries who choose an HMO option, that progress is necessarily limited by the substantial changes in organizational structure needed to move from a completely fee-for-service system to a completely capitated system. It may therefore be desirable to consider methods which limit the scope of capitation payments to physician services, or to various types of physician services, rather than necessarily requiring full capitation immediately.

The difficulty of inducing HMOs to accept 95 percent of the AAPCC as payment for a Medicare beneficiary's full range of service has also been a problem. Perhaps a different payment system would be more attractive to physicians.

There do exist some precedents for "partial physician capitation" (PPC). In the End Stage Renal Disease program in Medicare, nephrologists are capitated for the routine doctor services associated with management of a dialysis patient. They can bill fee-for-service for non-dialysis related services, or for services provided to the patient during hospitalization, so this capitation is only partial. And there have been concerns raised--as there always are with capitation--of whether the services actually being provided are enough to justify the current capitation rate. But the method seems administratively feasible for this set of chronically ill patients, and does not seem to have adversely affected quality. We have no comparative evidence, however, of its impact on volume and intensity of physician services or of complementary services.

According to Hillman's recent study (1987), 46 percent of all HMOs pay primary care physicians by capitation of the individual physician. The extent of the use of capitation varies across model types. Network-model HMOs use capitation most frequently (76%) of the time, while staff-model HMOs use capitation only 10% of the time. These payment schemes usually limit the capitation payment to services provided by the primary care physician himself or herself, but 40% of HMOs require primary care physicians to pay for outpatient laboratory tests directly out of their capitation payments. Thus, the obligation for primary care physicians to pay for ancillary tests out of what might otherwise become their own income is an incentive for them to reduce the volume of such services. In addition, there are frequently financial arrangements that put the capitated gatekeeper physician at risk for part of the cost of other services used by his panel of patients. For example, 54% of capitated primary care physicians in HMOs have a percentage of their capitation withheld in case their use of the actuarially-budgeted referral funds are too high. Some HMOs add-penalties for high referral costs beyond the loss of the withheld amount of money, including, for example, placing liens on future earnings, decreasing the amount of the capitation the following year, and excluding the physician from the program. These withholds and penalties beyond the loss of the withhold account are designed to reduce the volume of services ordered (including referrals to specialists, outpatient tests, and hospitalizations - although not all HMOs place all physicians at risk for all of these different funds). Furthermore, any positive surplus in these referral funds may be shared with physicians as bonuses, a further incentive for the physicians to limit the volume of services ordered.

The evidence on the effect of these methods of payment (compared to fee-for-service) on volume and intensity is limited, though the rationale for capitation plus risk-sharing clearly is to control the use of other medical services by the gatekeeper's panel. We think we know from the SAFECO experience that an important factor that contributes to the physician's behavior is the number of patients in his or her panel that are subject to capitation-based financial incentives.

The common use of case management in pre-paid practices may be an important reason for the reduced expenditures in many health maintenance organizations, although the opportunity for case management need not be limited to pre-paid practices with risk being assumed by physicians. Indeed, the opportunity for high cost cases being handled through case management

offers substantial opportunities for volume control to the Medicare program. Cases to be managed could be identified in several ways, such as individuals who are diagnosed with diseases generally associated with high cost; individuals who have already incurred high costs and trigger the case management program by exceeding a threshold of expenditures; and individuals who undergo a high cost procedure (such as coronary artery bypass grafting). The identification of the case manager will be important, and could include physicians, social workers, nurses, and other health professionals. A case management system could be instituted with incentives, although even without incentives, a case management program might be effective either by using the case manager as gatekeeper or through the case manager's effective recommendation of more efficient practice patterns and recruitment of community resources.

5.4.2.1. Conceptual Issues in Partial Physician Capitation

Suppose for the moment that a particular physician has been assigned responsibility for caring for a particular set of patients for a particular set of illnesses. This could be the primary care gatekeeper model, where the doctor manages care for a set of people for the full range of illnesses, illnesses which are usually acute rather than chronic. The capitation payment would always cover the range of services the primary care physician provides, but it might or might not include all or part of the cost of other services his or her patients use. Or it could involve capitating a specialist for care for a particular chronic condition, on the ESRD model described above. What are the incentives this arrangement provides?

The clearest financial incentive applies to the services that the capitated physician will or could provide himself. If his motivation is to maximize real profit, he does that by providing as few services as possible, consistent with the maintaining of the size of his panel at what he regards to be appropriate levels. This will probably lead to provision of fewer of his or her own services, as well as the other services that are complementary to them, than would be the case under fee-for-service, other things equal. To the extent that the gatekeeper must compete with other doctors, including fee-for-service doctors, to retain patients on his panel, and to the extent that reduced services will cause people to leave his panel, then the difference in service levels between capitation and fee-for-service will be less.

What about other medical services that are not complements? Absent any risk for the cost of substitute services, the profit-seeking physician under capitation will attempt to substitute such services for services for which he himself pays the money or time cost. For services for which the demand is independent of the use of his own services, and which do not affect the demand by patients to be members of a doctor's panel, he will be incentive neutral. He will consider only whether the service improves the patient's well-being, but will not consider the cost of the service.

Putting the doctor at risk for part of the cost of referred services does induce him to consider those costs, but there is no obvious way of guaranteeing that costs to Medicare and benefits to patients will be correctly

traded off. For other services which substitute for those that could be provided by the gatekeeper, there will be an incentive to increase substitution under capitation as compared to fee-for-service. That incentive can be offset by putting the gatekeeper at risk for some of the cost of substitute services, but then there will be an incentive to provide too little of such services. There is just no obvious way to calibrate the incentive to get the volume just right, nor is there any empirical evidence to indicate what the right stimulus is.

5.4.2.2. Risk Spreading and Incentives

Capitation as a method of controlling medical costs depends upon the scope of the capitation payment. If a comprehensive capitation payment is made to a physician (intended to cover all of the medical services needed by his patient enrollees), then it is definitively true that total expenditures will be affected only by the number of persons at risk, not by variations in the volume and intensity of the services provided. In such a setting, the incentives for the recipient or manager of the capitation payment to control costs are strong, since every dollar of cost reduces net income by one dollar. However, just the opposite is true if the capitation payment covers only the services of the primary care physician, and there are no controls on the use of ancillary services, referrals, and hospitalizations. In this case, capitated physicians would most probably substitute these other services instead of providing them himself. While capitating primary care physicians for primary care services only is, therefore, not likely to be guaranteed to control costs, comprehensive capitation turns the primary care physician essentially into an HMO, something this is surely too risky for individual doctors. Indeed, the extensive administrative infrastructure needed by an HMO to budget properly for different medical services, not to mention the license to assume insurance risk needed by HMOs, indicates that individual primary care physicians in the community are not able to accept such a responsibility.

Finally, capitation--with or without risk for the cost of referred services--puts the gatekeeper at risk for the event of a severe and expensive illness by one of his panel of patients. Some of this risk may be averaged over a number of capitated patients (Medicare or non-Medicare), but there is a possibility of a high "cost" patient taking up much of the gatekeeper's time and reducing his real net income. Most HMOs deal with this problem (87% by Hillman's survey) by protecting the capitated doctor from the very high cost due to outliers.

How serious is this risk pooling problem likely to be? If current volumes of services are taken as measures of current cost, the work of Mitchell et al., suggests that there is likely to be substantial heterogeneity across services or patients. However, Langwell and Pauly argue that, for reasonable values of variance due to random factors, many physicians who handle the bulk of Medicare patients would be able to pool away much of the purely random fluctuation over time. This suggests that what Mitchell et al., are finding must be the result of a non-random distribution of patients across doctors.

Mitchell et al., have pooled data up to the level of the diagnosis or condition, and continue to find wide variance in the average value of charges across physicians who see Medicare patients. (They have not calculated values for all Medicare patients the doctor sees.) Either some physicians treat patients with the same conditions in very different ways, or average severity varies across physicians in a nonrandom way, with some physicians consistently attracting patients with more serious conditions. Obviously, a resolution of the question of severity vs. patterns of practice is crucial in evaluating the efficiency (and the equity) of capitation. If some doctors treat patients of the same type more intensively than others, and it is that variation which is being measured, then it is true that imposition of uniform capitation would hurt the former and help the latter. It does not seem inequitable to redistribute real income away from doctors who practice in an especially lavish way -- if that is what high volume and intensity per patient represents.

Let us suppose, however, that it is variation in severity which accounts for the observed differences in volume and intensity. One thing to note is that applying capitation only to new patients would produce less "inequity," as long as new patients have or can be made to have a more random distribution of illness severity. If some doctors do nevertheless consistently attract patients of above- or below-average severity, then no amount of pooling can work effectively to produce an equitable solution unless severity can be built into the capitation payment. With a uniform capitation, even pooling a large number of physician recipients would still be expected to yield income differentials within that set, since the doctor who brings costly patients to the pool will surely be less well rewarded, in equilibrium, than the doctor whose patients are low cost.

The only solution to this problem is to develop, if possible, a way of adjusting physician capitation payments for the severity level of the doctor's enrolled population. This is exactly the same problem Medicare has faced in determining the AAPCC for HMOs: it is probably a more serious problem at the level of an individual physician. Without such an adjustment, even the doctor who accepts capitation will have an incentive to reject panel members whose cost he expects to be high and to strive after members whose cost he expects to be low. That is, creating one risk pool will dilute the risk of "preferred risk selection" (in the sense of some physicians attracting sicker patients) if the risk pool can distribute payments in varying amounts.

An important alternative strategy is based on the ESRD model. The difference here is that the capitation is not intended to cover all of an individual patient's needs comprehensively, but rather to cover only a subset of that patient's medical needs (i.e. services related to dialysis). Nephrologists are capitated only for the routine doctor services associated with dialysis, not other medical conditions that may arise. In essence, the capitation amount then is awarded to a site of care in order to take care of a patient for a specific set of services. Other "sites" or organizations that might form the basis for capitation in the same way include home care, rehabilitative care, and long-term care facilities. Doing so would, of course, be more difficult than in the ESRD program, since there is no specific condition or service to use to identify obviously homogenous groups of

patients. And yet, "Medicare beneficiaries receiving nursing home care" might, upon further study, be found to be not much more heterogeneous than "Medicare beneficiaries receiving outpatient dialysis," at least with regard to need for some subset of services. Again, the common theme among these sites is that the services required by patients are more predictable and routine than, for example, all the services that might be needed for any reason by a patient with rheumatoid arthritis.

It should also be noted that if preferred risk selection occurs, raising the capitated price for all patients will not necessarily solve the problem. With higher capitation fees current physician income could be maintained by reducing the number of patients in the panel. Thus, physicians may opt to skimp even more and search harder for lower risk patients than they did before. To ensure access for high risk patients, especially for the poor, a risk-adjusted fee needs to be developed; there is simply no other way to avoid the disincentive associated with money-losing patients. Alternatively, if special physicians seem to attract high risk patients, it may be possible to designate certain physicians under special contractual agreements to care for high risk patients.

Except for the minority of patients who might be capitated based on "site," the significant administrative difficulties, and the perverse incentives in partial physician capitation suggest that it may be less desirable than some other alternatives. Indeed, given the additional problems of defining what services are at risk and how payments should be divided, partial capitation probably is less attractive to the market and to Medicare than is full capitation.

5.4.2.3. Design Issues

One important design issue is whether capitation payment is to be voluntary on the parts of patients and/or doctors, mandatory, or subject to financial incentives. If physician acceptance of capitation is voluntary on a patient-by-patient basis, the issues of self-selection based on expected patient severity are magnified. An intermediate strategy would be to offer incentives for accepting this form of payment, as in the case of the participating physician program.

The third possibility is simply to switch all Medicare payment for certain disease or types of services to a capitation basis, perhaps after a voluntary phase-in period. There are precedents for such a strategy. Nephrologists providing routine physician services to ESRD patients were originally paid on a fee-for-service basis, then were permitted the option of capitation for a time, and eventually were all put on capitation. Medicaid programs have assigned some beneficiaries to capitated programs. Particularly if mandatory assignment is already imposed, switching physicians to a capitation basis appears to be administratively feasible if it is politically feasible.

Another important design issue is the determination of the capitated fees. Currently, the AAPCC-limited fee for HMO patients is 95% of the average

cost of patients under fee-for-service. This average includes a small portion of the beneficiaries who account for a substantial portion of expenditures under fee-for-service. In 1982, 5% of patients accounted for 54% of Medicare payments (Sisk et al., 1987). If a fair percentage of this minority is not treated by the capitated system then the rates paid to the capitated system would be too high. HCFA's total expenditures may, as a result, be higher than before capitation.

Given the incentives in a capitated system, what could be said about the effects on quality of care, particularly for the elderly population? There is little documented experience with provision of care by HMOs to the elderly or poor (Hammons et al., 1988). Capitated systems may have problems delivering care to the elderly who often present a variety of illnesses that require special methods for delivery of care. The model using a gatekeeper who bears financial risks beyond primary care may be inappropriate for the elderly because it restricts their ability to receive necessary specialized care.

It may therefore be worthwhile to consider a capitation system where only primary care services, especially office visits and consultations for referrals, are capitated, while other services are not capitated. Primary care services are also, in any case, difficult to monitor through utilization review. However, the nonprimary care services could be monitored through utilization review.) Such a partial capitation will not require as complex a bureaucracy as would full capitation. It could also be affected by exempting the visits to the capitated physician from out-of-pocket expenses but requiring out-of-pocket expenditures from the beneficiaries for referred services, or it could be the method used by a Medicare PPO for primary care services.

5.4.3. Summary

Although health services investigators have demonstrated that health maintenance organizations are capable of reducing expenditures, principally through a reduction in hospitalization and length of stay, HMOs have not been shown to moderate the rate of increase of health expenditures over time. The likelihood of capitation having a substantial impact on Medicare expenditures will depend in part upon the risk arrangement that is implemented and the organization of the capitation programs.

The reason for capitation having a salutary effect on volume is not clear. While some observers believe that physician incentives are important, others have suggested that the group process of physicians working together and the effect of capitation on the organization of practices are the principal influences. Therefore, it is possible that similar changes in practice organization could lead to different practice styles independent of a change in financing mechanisms. Case management may also play a role, though primarily in conjunction with financial controls.

Although capitation has the potential to control Part B volume and expenditures, a number of problems related to capitation make it less attractive as a volume control than other approaches. Organizational changes

necessary for a completely capitated system would be difficult to realize in the short term (and maybe even in the longer term). Capitation on a partial basis raises the possibility of increased referrals for services not included in the partial capitation, which could preclude significant cost savings. Either partial or complete capitation present risk pooling problems that require case mix severity adjustments; at present, these methods are not well developed. Likewise, either type of capitation requires an equitable determination of capitation fees. Given the current problems with AAPCC's for Medicare HMOs, a successful approach to setting appropriate and equitable fees seems unlikely in the near future.

Appendix: Theoretical Effects of
Capitation Under Behavioral Models

Appendix: Theoretical Effects of Capitation Under Behavioral Models

Capitation Under Profit Maximization

In this arrangement, the patient will choose a physician and notify HCFA or the carrier of his choice. The carrier in return will pay to the provider a fixed annual rate per registered patient, for a pre-determined set of services.

The most simplistic and straightforward analysis is to imagine that the physician is the provider who receives the annual fee. Once the patient is registered, the physician will maximize his profits if he or she provides the patient with zero services. However, if the patient did not receive any services even though he was sick during the past year, he may not register as the physician's patient in the next year. The patient will be better off taking the money himself and purchasing health care in the open market (if that option is made possible). Alternatively, the patient will look for physicians who are ready to provide some health care.

Will the patient find physicians who are ready to provide more than zero health care? The plausible answer is yes, if there is excess supply at the capitated price. Competition among physicians for enrollees will provide incentives to provide some services. Physicians with low enrollments will have an incentive to increase enrollment. They will attempt to attract patients by providing some service for a smaller profit than the maximum profit. The level of the services provided will depend on the ratio of patients to physicians in the market area. Since the physician has an incentive to maintain the size of his panel and even to increase it, he will provide care at a quality level sufficient to maintain his reputation and avoid obvious disasters. The level of quality will also depend on the patient's ability to monitor the care provided. External quality monitoring by the paying agency will also be necessary. Noncompliance with minimal quality standards could result in exclusion from the physician panel.

Under capitation, financial risk for expenditures beyond the capitated fees is imposed on the physician. This risk can be pooled by physician groups or by developing fee schedules which correspond to the level of risk a patient represents.

In a capitated system, a profit maximizing physician will attempt to maximize his revenue by attracting a large number of patients to his panel and minimize his cost by providing those patients with the minimal possible amount of services.

The above strategy is likely to result in the reduction of volume of services, compared to fee-for-service. Level of expenditure will be a function of the number of beneficiaries and the determined capitated rate. The expenditures will be fixed and predictable and could be calculated to be smaller than the expenditures under FFS.

Access and quality of care will depend on the relative strength of two offsetting incentives. The incentive to maintain panel size will contribute

to improved access and quality while the incentive to reduce costs will contribute to a reduction in services per patient and to the adoption of cost cutting measures which may decrease quality of care. Quality is particularly likely to be affected by the use of low cost inputs and denial of necessary services. Thus, services that were effective under fee-for-service may become ineffective under capitation if input substitution takes place, or if those services are provided without complementary services that were provided under FFS.

A private market for "better quality care" may develop outside the "official" capitated system. Thus, beneficiaries' burden of the costs of care may increase. Moreover, such a phenomenon will result in unequal access and inequality in the type of care received.

Capitation Under Target Income

As in the profit maximizing model, we assume that physicians are directly capitated and that risk pooling arrangements have been created that ensure physician participation in the system.

Physicians' behavioral responses to the shift to a capitated system is determined in this model by income and the subjective valuation of the professional services they provide. To attain the level of desired income, physicians need to determine the size of their patient panels. Increases in the panels' size increase revenues. It also increases work since subjective standards of professional behavior require that each patient will receive, on the average, a given amount of care. In determining the revenue, the income effect creates incentives to increase the panel's size. The substitution effect creates incentives to provide less care as leisure becomes scarce and its marginal value increases. The substitution effect creates incentives to either reduce panel size or provide a lower quality of care.

The incentive to reduce the quality of care is reinforced by the incentive to reduce costs. This can be obtained by reducing the costs of inputs and reducing the provision of complementary services whenever possible. These incentives must be balanced with the physician's subjective valuation of his or her professional conduct. Thus, in this model the quality of care is determined by the physician's subjective judgement and the need to ensure the size of the patient panel.

The model has the following implications:

1. Volume of services will depend on the prevailing market conditions. Excess supply will result in increased volume in order to attract patients.
2. When markets are characterized by excess demand, at prevailing Medicare capitated price, the likelihood of low quality care will increase.

3. Access to care may be restricted by skimming. This is particularly likely if excess demand prevails. The physician's subjective valuation may be better rewarded by providing good care to his panel patients, with whom he develops a personal relationship, than from accepting new patients, which will imply more work and a reduction in quality of care for everyone.
4. Increase in capitation rates to resolve access problems may actually result in reduced access. The higher rate means that a fixed income could be obtained by seeing fewer patients and providing better care to those remaining in the panel. The use of medical benefit organizations transfers the skimming function to the organization and away from the physician and makes skimming more likely. An alternative reaction to increases in the capitated price is to provide care to more patients and let each patient consume more resources, especially more referrals. Whether this reaction occurs depends on the contractual arrangements regarding services not provided directly by the physician.

Capitation Under Patient Agency

In the agency model, the direct financial influences of capitation on the physician are not important, and may be no more effective (from the perspective of this model) than posting utilization results in the physicians' lounge. However, the effect of capitation may be substantial by way of its effect on the organization of practice, on professional relationships, and on relationships with patients. Reorganization of practice is often forced by capitation, for instance, to take advantage of economies of scale and thus provide care more efficiently, or to share risk across more physicians. The way in which this reorganization takes place may either increase or decrease the volume of services. In general, however, these organizational effects will probably decrease utilization, even independent of the financial effects.

5.5. Expenditure Growth Targets

5.5.1. Overview of the Literature

Expenditure targets have not been utilized in this country, thus there is no empirical data assessing effects on volume or expenditures on the U.S. population. A description of the expenditure target system used in West Germany can be found in Schulenburg, 1983, and of those used in Ontario and Quebec in Leader et al, 1988.

5.5.2. Likely Effects of Implementation

Ever since the surprising jump in Medicare Part B payments in recent years, there has been more intense interest in developing physician payment systems that set expenditure, reimbursement, volume, and intensity growth at appropriate levels, and that permit Medicare to budget both its total expenditures and the Part B premium beneficiaries pay. There has, however, been a fundamental difference of opinion among many about the purpose of payment reform. While some see payment reform as a program for cost-containment, others believe it is a program to achieve broader goals including: improving the quality of care, the appropriateness of services, and the access to care (both short-term in maintaining the willingness of physicians to treat Medicare beneficiaries and long-term in its effects on health manpower), as well as controlling costs to the patient. Inter-physician equity has also been discussed, but primarily as a means to the other goals.

Expenditure growth targets offer the advantage of an explicit budgetary target. This is in contrast to the more reactive budgeting that has been done in recent years through the budget reconciliation acts of Congress. The greater predictability that this method can provide is an advantage to all concerned, including HCFA, providers and beneficiaries, and is one of the advantages of this approach to controlling volume and cost. Expenditure growth targets, by definition, concentrate first on the objective of cost control; it represents a strategy of pursuing cost control explicitly, but subject to the constraint that quality, access, appropriateness or equity not be adversely affected. It is surely impossible to find a scheme that guarantees success in achieving all goals simultaneously.

Why should one presume that growth in Part B expenditures, no matter how great, is necessarily a problem? Growing cost is a problem for the beneficiary premiums and general tax financing that must cover the increased cost, but if one felt that the benefit to the elderly--in terms of health, comfort, or convenience--were great enough, one might applaud the cost increases that accompany these gains. The real problem is the absence of evidence that the extra costs really do go for things that are "worth it." Why should there be waste?

Most thinking on this matter to date has looked at the problem as one of incentives (usually assumed to be "distorted" in some way) to physicians and, to a much lesser extent, to patients. The objective has been to alter those

incentives in a way which, it is hoped, will yield lower or more appropriate expenditure and resource growth. The problem with this approach, as shown elsewhere in this report, is that in theory and in practice incentives have at best, ambiguous results, and at worst, can lead to major errors because of the difficulty of controlling or predicting the response to them. In some cases, (e.g., copayments or capitation), we can predict the effect on Medicare cost, but we always have difficulty predicting the effect on appropriateness of care. (Indeed, we have difficulty defining "appropriateness.") Incentives are blunt instruments, hard to fine tune for precision work. One cannot even be sure whether they will increase or decrease volume and costs.

The use of expenditure targets as a volume and cost control measure represents a somewhat different strategy. By first defining appropriate aggregate Part B expenditure (reimbursement) growth, expenditure caps would point out deviations from the target level of expenditure growth that are due to unprojected and undesired changes in volume. When such deviations are detected, the target would then set into motion changes in unit prices which will (with a possible brief lag) set expenditures at the correct level. Here we consider expenditure targets both with and without balance billing allowed.

5.5.2.1. Elements of the Expenditure Growth Target Scheme

We begin by describing the "no-balance billing" version of this mechanism. It is easiest (though not essential) to define the mechanism for a particular population. There are three essential elements in the payment mechanism:

- (1) An expenditure determination model,
- (2) a unit price adjustment formula, and
- (3) a timing choice.

The expenditure determination model would determine how expenditures ought optimally to grow for the population in question, taking into account changes in input prices, desired volume of technical change, and demographic or health changes in the population. Deviations of actual from optimal expenditures would then cause offsetting adjustments in unit prices. In the simplest "one-for-one" adjustment formula, increases in expenditures in excess of the optimal level cause reductions in unit price intended to be sufficient to hold total expenditures at the optimal level. It would also be possible to have other "strengths" of adjustment. For instance, half or some other fraction of excess expenditures due to volume could be captured through the unit price adjustment. There is also a question of whether the adjustment should be one-sided, reducing price for excess volume only, or two-sided, raising price as well for lower than optimal expenditure growth.

On the one hand, a two-sided adjustment would be more equitable if physicians are able to maintain an acceptable quality of care while becoming more cost-efficient and saving the system money. On the other hand, a two-sided adjustment might provide a strong incentive for underprovision of

services. In part, this decision depends on the amount of confidence placed in the accuracy and appropriateness of the target, as discussed in "Sharing the Cost of Overruns," below.

The third element of an expenditure target scheme is the timing of price adjustments for excess expenditures. Two variations are:

1. Withhold a percentage of payments until the end of the time period. If there are excess expenditures, distribute only the amount that keeps the total expenditure within the budget total.
2. If the target is not met, readjust payment rates for next period so that, given projected volume, excess expenditures (or a portion of them) will be considered in the calculations of unit prices for next period.

The first method is frequently used by HMOs to pay primary care gatekeeper physicians but has not been used for a total physician budget. It also approximates the scheme used by sickness funds in West Germany. The second method has not been used in the United States.

Use of expenditure growth targets does not preclude the use of other volume controls, such as strong utilization review, case management, guidelines/education, and other utilization management tools. In fact, it gives the medical profession a motivating rationale to develop, support and even implement utilization management.

In what follows we explain in more detail how such an expenditure-limitation system might be designed, and how it might be expected to work.

5.5.2.2. Projecting Expenditures

In order to use a system which limits physician payments, a method must be developed to estimate the desired target level of expenditures. Ordinarily ideal expenditures for the future will need to be forecasted or projected, but it is surely possible to make after-the-fact adjustments when unforeseen changes in outside influences occur. This is the level which, if achieved, will yield a revenue to physicians just equal to the amount billed. To a considerable extent, the problem here is similar to that in making or rationalizing any budgetary estimate, such as setting DRG payment levels or projecting average adjusted per capita expenditures for purposes of HMO reimbursement. The main difference from DRG pricing, where prices per unit of service are of primary concern, is that technological changes, input price changes, and changes in the characteristics of the population will affect the intensity of services provided over an episode of care as well as the cost per unit of service. This is not, however, as different as it seems, since one can view a hospital admission that is reimbursed via DRG payment as a composite of services that can vary in intensity. In any event, the need to adjust for technology is the same. The main difference from the AAPCC is that there is no average Medicare benchmark. Here again, the population demographic adjustments are similar.

Several factors related to technological change complicate the development of an expenditure determination formula. First, if the expenditure target is for Part B only, unforeseen technological changes that shift services from the inpatient to the outpatient setting will cause short term problems by causing Part B expenditures to exceed the target, while overall program costs remain stable. In the longer run, this "penalty" on Part B costs may serve to discourage the development of useful but expensive outpatient technologies, or impede appropriate shifts from the inpatient setting. However, to the extent that future technology can be foreseen, cost-effective shifts to Part B can be built into the targets. Second, some mechanism is needed to deal with the impact of new clinical standards that have an impact on utilization. For instance, if authoritative standards were issued that called for yearly mammography for women ages 65 and above, physicians would be faced with a dilemma: should they provide this standard of care, knowing that they will be penalized financially when Part B costs exceed the target? Building such desirable or prescribed changes into the target would be necessary. Third, an assurance that new benefits mandated by Congress would be fairly budgeted would be required for acceptance of such a control by physicians.

Some hard decisions would need to be made in order to arrive at a tolerable formula. In effect, the formula should incorporate all the things that are regarded as "legitimate" reasons for Part B expenditures to increase, and should quantify the appropriate amount for them to increase. Demographic variables--total number of eligibles, age, marital status--seem obviously legitimate candidates, and their potential impact on volume could easily be determined from statistical analysis of relative volumes. The impact of varying illness levels could be quantified in much the same way. Probably the most difficult task will be to judge the appropriate change in technology, and the impact it should have on total expenditures. Explicit decisions about "leader" technologies will have to be made. Likewise, adjustments for productivity changes will be controversial. Finally, adjustment for input price changes should be based on the Medicare economic index of input prices.

As with DRGs, it seems reasonable to calculate different projected expenditures for populations in different areas, but the designation of appropriate geographic areas, as in the case of hospital payment, is also likely to involve some controversy. Other disaggregations are, in principle, possible and there are several principles that should guide these. First, it must be kept in mind that the expenditures are for beneficiaries, not physicians; thus, a population-based approach is needed. Second, the population covered should be large enough to be administratively feasible and to avoid a large number of small administrative units. Third, the physician population within a target area should be small enough to allow for peer interaction and influence. Fourth, it would be ideal to use current structures, such as existing peer review organizations, to administer the target; while the objective here would be to avoid development of new organizations, that may not always be possible and should not be completely ruled out. Finally, the size of the area should be one that stabilizes year to year variations in expenditures, so as to make the targets predictable. One could, however, imagine separate expenditure targets for physicians in

different specialties, or even for the set of physicians on a hospital's staff if the hospital serves a large enough population of patients.

There are two design questions here; (1) What population should have its experience aggregated to define an expenditure target? (2) What services should be combined to determine the target and to determine the degree to which the target was achieved?

The first question addresses in part the usual tradeoff between incentives, selection, and the risk-reduction benefits of pooling. Suppose a target was defined for all physician expenditures in a county. For all but very rural areas, any random or unmeasurable occurrence of illness would be pooled or averaged enough to dampen any impact of random illnesses. Measurable systematic changes in illness, such as epidemics, could be built into an after-the-fact adjustment in the target, as would any unforecasted demographic changes. Important changes in the extent of patient "border crossing" could be detected on a sampling basis; there would be no need to record county of residence for all patients. Pooling over an area as large as a county would also be sufficient to prevent any selection of patients expected to have lower or slower-growing medical care costs. But the pooling that disperses risk also diffuses individual incentives. No individual doctor or small aggregation of doctors will gain directly from controlling volume.

If the number of beneficiaries whose experience is added up to determine whether the target was achieved is small, the number of doctors who treat them may be sufficiently small to offer incentives for, or permit easier coordination of, attempts to limit volume. The cost of better incentives is that the group of doctors will be vulnerable to volume effects of random serious illnesses, and may also be motivated to accept or reject patients for treatment based on the impact they are expected to have on volume. If the areas are defined narrowly, there may also be a greater possibility that one doctor will have patients operating under more than one expenditure cap (if his or her "market area" overlaps two different target areas), which will be confusing. Finally, the administrative difficulties of computing and monitoring many targets will be greater, although a formula approach might ease matters here.

On balance, there seem to be considerable risks to making the mandatory geographic areas too small. An area large enough for risk pooling and economical administration will be too small for incentives to matter in any case. What might be envisioned is a strategy in which targets are defined for an area such as a county, but in which doctors serving a single population could voluntarily elect to have a separate target defined and monitored for "their" population. Whether beneficiaries would be allowed to have a say in this matter is an open question, but an equally serious question would be whether preferred risk selection could be controlled. If it could--if the population can be defined so that it is stable, and not especially selected to be likely to have a lower rate of growth in cost--then such self-designated groups would be permitted. Note that if populations are defined by the geographic areas in which they live, and targets are defined by a rate of growth from some base level, the HMO problem of preferred risk selection will

not arise, unless healthier populations would also be expected to have a lower rate of growth in volume from their lower base.

The other design question is which services should be aggregated. Ignoring administrative complexities, there is no reason not to set different targets for different services, and measure the achievement of such targets separately. We are thinking here of basing aggregation on type of service, not self-designated physician specialty. If uniform targets were set for all services, but if only some services overshot the target, then it would be sensible to adjust prices only for those excessive services. Doing so would also reduce the "inequitable spillover" problem. It might be desirable as well to set different targets for different services. One might target a lower (or even negative) rate of growth in volume for surgery or some other service believed to be in excessive volume, and yet permit expansion of volume and expenditure for patient visits. This would be an indirect way of adjusting relative prices closer to an RBRVS or some other relative price schedule, but which would automatically readjust prices for whatever volume impacts would occur. Given the uncertain nature of volume impacts of RBRVS, or any other change in relative prices, such a fail-safe limitation on total expenditure might be useful insurance.

Once a formula is developed to determine optimal expenditures for a given population for a given time period, it will be desirable to monitor the performance of actual expenditures against the target. If expenditures begin to exceed the target, the first question to be addressed is whether there is a good reason for this happening. An epidemic, an unexpected surge in malpractice premiums, or the delivery of a highly beneficial new technology could all be reasons for permitting upward growth in spending in excess of the initial target. In this sense, good reasons for expenditures to grow can easily be incorporated into after-the-fact adjustments, so that such a system can deal quite effectively with such things as epidemics as long as they can be objectively identified.

5.5.2.3. Sharing the Cost of Overruns

A perfectly fixed expenditure target would imply that physicians collectively would share in the cost of overruns and underruns dollar-for-dollar; unit prices would fall in overruns but increase in underruns. But things could be made more flexible. Since there may be some uncertainty about whether additional costs are legitimate or not, one could imagine that the Medicare program would agree beforehand to accept some fraction of the overrun. Or there could be a corridor (a percentage of total costs or a fixed dollar amount) of full physician responsibility for overruns, followed by partial Medicare sharing. The size of this corridor could be based on the degree of predictability thought to be obtainable with the formula used, and the details of such an arrangement could be varied to put more or less of the burden of overruns onto physicians or onto Medicare.

It may also be possible to quantify the confidence one has in the expenditure determination for any time period. The adjustment for appropriate technology, in particular, may be more or less certain, depending on the

nature and number of new technologies available. In times of uncertainty, it would be appropriate for physicians and Medicare to share the cost of overruns.

Another design question is whether underruns--actual expenditures less than projected--should prompt a "dividend" to physicians in the form of higher unit prices. The underrun question is subtle (though perhaps not especially likely). Why might expenditures undershoot the target? If the reason is that sometimes there can be good exogenous reasons for spending to grow less than forecasted, then there is no need to share the value of this deviation with doctors. If, however, such underruns actually represent underservice, one could argue for using such an event as a trigger for paying doctors more, if higher fee levels will stimulate more service. There is, as we have already noted, some doubt about whether volume responds to price in this way, and there may be even more doubt about whether price-increase-induced volume growth will occur when and where the underservice occurs. A sensible design strategy, at this point, might be to make the unit price adjustment symmetric, but to monitor carefully to diagnose the causes of underruns should any occur.

5.5.2.4. Copayments and Expenditure Caps

Under the "withhold" model of setting expenditure targets, there would be some ambiguity in determining the copayment for beneficiaries since the final price is not known until total volume is measured. One possible strategy would be to base the copayment on what the unit price would be if volume targets were met, since consumers can hardly be expected to forecast physician volume and adjust their demand accordingly. In the "next period adjustment" model, an overrun in one period would lead to lower copayments the next period, which would tend to stimulate demand rather than discourage volume. In this sense, as usual, demand side incentives and supply-side incentives move in opposite directions. If these lower copayments were felt to be a serious enough problem, it would be possible to base the next period's copayments on what unit prices would have been had there been no adjustment for previous-period overruns.

5.5.2.5. The Timing of Adjustments

As noted above, Medicare could either "collect" for overruns from an amount withheld based on initial nominal payment rates, or it could adjust prices in the next time period--given forecasted volume for that period--to recover the overrun. The two strategies have different implications for physician behavior.

The main difference is that, under the withhold strategy, the distribution of the reduction is made on the basis of a doctor's past volume, which cannot be changed, whereas under the "next period" model, recovering the overrun will alter the price the doctor faces in the next period, and give him or her the opportunity to adjust volume to that new price. That is, under the "next period" model, the price in that period will necessarily be lower if there is an overrun in the previous period.

An example will illustrate. Suppose doctors expect a 10 percent overrun in period 1, and no overrun in period 2. Suppose also that the nominal price (NP) is to be the same in both periods. Then the expected price that an individual doctor will face under the withhold strategy will be a .9 NP in period 1 and 1.0 NP in period 2. In contrast, under the "next period" strategy the individual doctor would imagine the net price to be 1.0 in period 1. But he or she would also expect the price for additional volume to be .9 in the next period. That is, if every other doctor provided the forecasted volume in the next period and he or she contemplated providing one more unit, the marginal revenue from that unit would be .9. To the extent that volume responds to marginal price--a topic to be discussed below--the two schemes could have quite different effects on volume in any time period.

There is an intermediate strategy. Any overruns in period 1 could be collected in period 2 by "taxing" each doctor's period 2 payments based in period 1 volumes. This would be a lump sum tax equal to the doctor's proportionate share of the period 1 overrun. Only if a doctor's period 2 billings were less than the tax would there be problems.

5.5.2.6. Effectiveness at Limiting Volume Growth

Here we enquire whether this device is likely to be effective in controlling growth in expenditure and volume. We also examine whether it will be an appropriate control and whether the scheme is likely to be regarded as fair and feasible.

There seems to be little reason to doubt the effectiveness of expenditure limit or expenditure budgeting schemes in controlling Medicare expenditures. Whatever volume doctors select, the process automatically adjusts payment per unit so that total expenditures reach the target. The only way to frustrate this control in the withhold strategy would be to increase volume so much that all of the funds withheld would be used up. In the "next period" strategy, volume could conceivably be increased in the next period, but then there would be an even larger reduction in the period following that. A degenerative process in which price goes to zero and volume goes to infinity is theoretically possible but obviously implausible. At some sufficiently low price, willingness to provide large volumes of services to Medicare patients must diminish.

5.5.2.7. Expenditure Targets and 'Beggar My Neighbor'

If the volume needed to meet an expenditure target at some fee for service price level is the volume doctors will in fact choose to supply, then the target will be met exactly, the system will be in equilibrium, and total expenditures will be at the proper level. But suppose that the volume at the initial price level is expected by most doctors to be greater than the volume consistent with the target. This means that net prices will be reduced, and a doctor may be motivated to increase his or her volume in order to keep his or her income up. Robert Evans (1988) has called this an example of "beggar my neighbor" policy, as each doctor tries to snatch a larger piece of the total

expenditure "pie." He is concerned that the price decline that would follow such a strategy, especially if targets are set on a national or large geographical area level, would lead to prices which are not "credible." How serious should we expect such motivation to be?

We need to begin by noting that, since the existence of target income behavior has not been definitively established, the conclusion that the doctor will want to provide more services at a lower unit price, and that he will be able to do so, is far from certain. What we can say is that the process cannot spiral down to zero, since supply at a zero price is surely zero (except for voluntary labor by physicians offering free care to some Medicare beneficiaries), especially if the price is a Medicare-only price and doctors have the opportunity to treat non-Medicare patients. In a world of foresighted physicians, as noted below, the price and volume will immediately collapse to as low a level of price as is needed to reach the target level of expenditure. Will that price level be "(in)credible"? That is, will it be very low?

The answer really depends on something we do not know: how far must price fall before the target income motivation to generate demand--if it exists--ceases to have enough of an effect to hold expenditures up. Volume can still increase, but as long as it does not increase more rapidly than price falls, expenditures will still fall. In the West German expenditure targeting system, prices have surely not collapsed.

More generally, after several iterations, it is likely that doctors would learn their lesson--that volume increases are self-defeating. There would probably evolve (as game theory studies of this issue sometimes show) a "learned parallelism" of strategies that will prevent doctors from offering unbelievably high access and quantity to Medicare beneficiaries.

In short, expenditure controls would be virtually guaranteed to be a way of controlling Part B spending and premiums. This process would introduce an element of certainty into government budgeting and beneficiary expectations of future premiums.

While expenditure targets do not have the direct appeal of idealized enforcement of ideal clinical guidelines, they do represent an option which cannot fail to control expenditures no matter what the circumstances. That is, their effect will be less specific, less fine tuned, but more certain. This greater certainty should appeal to Medicare and to physicians, compared to the uncertainty about whether other methods will work at all, which will do more good than harm, and which will be subject to uncertain intrusion into doctor-patient decisions. Compared to utilization management or even organized managed care, an expenditure target -- since it only needs to affect unit price -- leaves the doctor much freer to practice medicine in the way he or she regards as best.

It is less certain, however, what effect this scheme will have on actual volume and on the appropriateness of the services provided (i.e. decreased overall volume will probably lead to decreases in appropriate, as well as inappropriate, services). Obviously, the volume of services can increase in

an expenditure limitation program even if total expenditures are constant, since decreased unit prices will occur if there are overruns.

But if physician services are substitutes or complements for other Medicare-covered services (including Part A spending), there can still be consequences for total Medicare spending. Suppose, for instance, that some physician services are complements for other Medicare services. More doctor visits may mean more drug prescriptions, more lab tests, or even more episodes of hospitalization. If budget overruns cause increased volume in Part A as well as in Part B, total Medicare costs could be adversely affected. The critical question then is whether the prospect of reimbursement limits would cause physicians, individually or collectively, to change overall volume.

Let us first consider a case in which doctors make volume decisions individually. How we describe behavior depends very much on the behavioral model which we assume to be appropriate, a subject we discuss in other parts of this report. Whatever the motivation, here the physician is assumed to consider the economic consequences of his or her action. What will be relevant for financially motivated decisions about volume is the "expected additional revenue" from providing another service.

The best benchmark case is one in which all doctors expect that there will not be an overrun in the current period. Then the nominal or posted price is the expected additional revenue from any service. If that price is consistent with equality between the desired physician volume and the forecasted physician volume, then the expectations will be realized, and the forecast will be met.

Now consider as an alternative the case in which an overrun is expected. This can happen when the volume doctors expect to produce in the aggregate at the nominal price is greater than the volume assumed in the forecast. Then the expected additional revenue is the nominal price less the expected reduction in price when the overrun occurs. If physicians and patients respond to lower prices by increasing volume, as many suspect, then there is a likelihood of aggregate volume increases-even if expenditures will not increase. Of course, the process is critically dependent on how physicians form expectations. If they envision overruns at the volumes that would accompany the initial (high) nominal price, but then understand that price would be cut, they may go on to forecast a yet greater increase in overruns, and a yet further after-the-fact cut in prices, driving the system to the "zero price-infinite volume" point described earlier. The result also depends on the substitutability of inpatient for outpatient care: if physicians can easily substitute hospital care for an outpatient service when overruns are expected, they will do so, thereby increasing overall program costs. But it may be more reasonable to assume that expectations are taken in stages, so that this will not happen; that is, it may be reasonable to assume that physicians will not look beyond the first stage.

How the process will actually work depends on how well doctors can forecast what total volume will be. If they can forecast accurately, the process of adjustment can be shortened. A numerical example will illustrate. Suppose the price for next period is 100. If, at that price, the actual

volume is expected to be equal to the volume consistent with the target expenditure level, the target will be realized. But suppose that all doctors correctly expect that, at a price of 100, actual volume will exceed target volume by 10%. Then the price would fall to 90. What happens next depends on forecasts of volume at a price of 90. If, at that lower price, volume is the same as it would have been under a price of 100, then that volume and a price of 90 will occur. In contrast, if the volume will be different at a lower price, then 90 is not the correct price forecast either. If volume will be higher at 90 than at 100, the price of 90 will not represent a price which meets the volume target; the price will have to be lower. So price will eventually have to be the price at which forecasted volume and the volume consistent with the expenditure target are equal. Thus expected (and realized) actual price will automatically "jump" to exactly the level needed to reach the expenditure target. A similar story holds for volume decline as a response to a lower price, as long as the price adjustment is symmetric.

In short, with perfect forecasting any expenditure target becomes automatically a self-fulfilling prophesy. It will not even require a period of ratcheting to the correct equilibrium.

Financially motivated physician behavior may well not be the whole story. Setting an expenditure target makes every physician's income depend on the volume decisions of every other physician. For the group as a whole, volume increases do not lead to increases in revenues. Volume changes only serve to alter the shares of a pie of fixed size. Recognizing the financially damaging effect of a volume increase strategy, physicians collectively may well take steps to control volume--or even reduce it if price is increased with underruns. The main issue here is one raised in our discussion of behavioral models. If an overrun does occur and unit price is cut, will that price cut stimulate further volume increases? It is obvious that a process of falling unit prices and rising volumes cannot continue forever. There is also considerable empirical uncertainty about whether volume really does always rise when prices fall. Finally, if a given price level should accompany an appropriate volume level, then there would be no need to alter the price, and the price level would be self-sustaining. There are many forms this collective behavior control could take, and its form will depend in part on the size of the set of beneficiaries for whom targets are calculated. It is unlikely, for instance, that all, or even a majority, of physicians across the country could organize a collective effort to regulate the volume of services provided nationally. However, it is conceivable that all physicians in a given market could develop mechanisms to monitor themselves and sanction those physicians who exceed reasonable volumes.

To be sure, for a number of physicians held to a given target, each individual physician will reasonably ignore the effect that his or her own volume changes will have on the unit price. But the group of all physicians cannot ignore this effect. The smaller the group, the easier it should be to organize collective action to do something about it. Actions in part may be "moral persuasion", encouraging other physicians to be conservative in therapy and to avoid services of low marginal benefit. It is not inconceivable that some professional self-regulation of volume could emerge in this setting. It is important to note, however, that antitrust issues are likely to arise with

any significant sanctions against outlier physicians; resolution of these antitrust issues would therefore be critical to effective self-regulation within the profession. For example, anti-trust regulation may be necessary to ensure that physicians do not erect additional barriers to new physicians seeking to establish practices and to share in the targeted expenditures.

5.5.2.8. Appropriateness of Volume Limitations

It is one thing to say that a device will limit expenditures; it is another thing to assert that the services that will then be provided will be the most appropriate ones. Even the recent substantial increase in Part B spending is only suspected to contain inappropriate volume, as Wennberg's cross-sectional studies suggest (1984).

The most obvious need is for some more direct way of determining and monitoring appropriate volumes to accompany any indirect financial device like expenditure limitation. Medicare needs to say what it wants to buy before it can conclude that any level of increasing expenditures is inappropriate. The obligation to determine a total budget pushes in this direction, but simply setting expenditures at a level sufficient to buy the desired mix and quantity of services does not in itself ensure that those services will be furnished. Someone will have to be equipped with the power and the information to judge appropriateness. That "someone" could be some type of review organization, or it could be Medicare beneficiaries themselves--with Medicare-provided advice on what represents appropriate services, beneficiaries may be able to come to a judgment about which providers are behaving inappropriately, and thus harness the direct effect of losing customers to whatever administrative sanctions the program may impose.

5.5.2.9. Fairness and Feasibility

The most obvious objection to an expenditure limitation scheme is precisely that it does hold all physicians responsible for the financial consequences of the volume decisions of any one physician. This means, to put it bluntly, that a physician who is making entirely appropriate choices can be punished, in the form of lower prices, for volume decisions of other physicians with which he is grouped. Since the Medicare volume of any one physician is too small to be a reasonable target for budgeting, this spillover is unavoidable. But it is also the incentive for the collective action which may well be the most helpful (and so far unused) way of controlling expenditures.

In one sense, expenditure limitation is like a scheme of global capitation for total payment, but with the division of the capitated revenues still based on fee-for-service. The experience of HMOs which have used such schemes would be relevant here, but is not generally known. What is known is that such withhold-adjust schemes are technically feasible, at least in small self-selected groups of physicians.

A final design question is whether it would be possible to have an expenditure limitation scheme implemented on a voluntary basis, with physicians agreeing to accept payment in this form, in return for a somewhat higher level of unit price. All of the problems of determining appropriate capitation payments would be present here, but in our view such a voluntary scheme, on a pilot basis, would be appropriate to explore. Related to this is whether special physician/patient groups such as HMOs or PPOs should be allowed to have a cap of their own. This might be desirable if these groups have already shown themselves to be efficient providers. It does, however, raise issues of skimming that might lead to inequity among providers.

One of the important design questions is how much Medicare should intervene in order to set such volume controls in place. In the Quebec system, there are elaborate--and constraining--limits on the income of each doctor and on changes in volume. West German sickness funds, in contrast, impose no special controls on volume, judging that getting more services for their members for less money can hardly be a bad thing. By analogy with the prospective payment system, it may be appropriate for HCFA to limit itself to setting the appropriate target, and the appropriate unit prices, and then leave it up to providers to decide how to react. Trying to impose a set of devices to "help" doctors cope can be well-intentioned but counter-productive. Since the fee level eventually gets reduced enough to induce a volume level which meets the target, perhaps HCFA can rely on this market-like arrangement without extensive further interference.

Given the likelihood that different doctors will respond differently to a volume control, it would probably be better to avoid complex administrative structures, at least initially. Subsets of doctors who seek to have volume limits calculated specifically for their members should be permitted, especially if these subsets can provide evidence of organizational changes intended to control volume.

5.5.2.10. Phasing and Voluntariness

The simplest way to look at revenue limits is as a more systematic form of past Medicare policy. But rather than try to approximate a target rate of growth in Part B spending by successive rounds of freezes, thaws, and relative decontrol, the revenue limitation strategy substitutes for such ad hoc policies an explicit process of determining what level of increase is appropriate, and of systematically adjusting average unit prices to achieve it. Since it may take some time to develop reliable methods to fix the appropriate rate of growth in expenditure, it may be desirable to phase in revenue limits by means of a "partial offset" formula discussed earlier. If expenditures exceed some provisional target, that could be the basis for trimming some amount off the increase in Medicare fees in the future.

The other design issue is whether revenue limitation, in whatever degree of stringency, should be compulsory for all doctors accepting Medicare patients. If it is compulsory, an unavoidable consequence is that some doctors may cease to take Medicare patients. An alternative would be to permit acceptance of volume limits to be voluntary, presumably for all

Medicare patients. If the incentive can be made "two-sided," the prospect of higher unit prices if volumes can be cut could even furnish a positive incentive to accept such a limit. But such schemes will only have partial effects. If Medicare is serious about controlling its expenditures, it will have to impose some kind of universal revenue limit.

The discussion thus far has been based on the assumption that no balance billing is allowed with the expenditure cap. In the following section, we explore the implications of an expenditure cap with balance billing allowed.

5.5.2.11. Expenditure Limitation with Balance Billing

How could a Medicare expenditure limitation scheme be combined with balance billing? In this case, the policy would limit Medicare's expenditure or reimbursement, but would not necessarily limit total expenditures on Part B services. However, the MAAC system limits total charges.

The formal structure of such a program would be much the same as that already discussed. An appropriate rate of growth in total Medicare spending, at current nominal Medicare payment rates, would be projected. If there was an overrun, unit prices would be cut, either by drawing down a withheld amount or by reducing unit payments next time period. For participating physicians, this would translate into lower prices for beneficiaries, and lower copayments. Physicians could also, as now, agree to accept assignment on a patient-by-patient basis; if assignment is accepted, the (lower) Medicare unit payment would be accepted as payment in full, except for patient coinsurance.

For a physician who does not accept assignment, the lower price could be offset by increases in prices charged to patients. Just as at present (with the exception of periods when charges to patients were frozen or limited), the physician could, within MAAC limits, offset what he or she regarded as insufficient Medicare payments by charging the patient more. No patient would be required to patronize a non-participating physician or one who did not accept assignment, but if the patient did choose to do so, he or she would be responsible for both the coinsurance and any excess or "balance" bill. A MAAC-like mechanism could be used here for control.

What would be the consequences of these variations in amounts patients are charged? The effect of lower coinsurance for services provided on assignment or by participating physicians would be expected to be an increase in the demand for physician services, which have become cheaper. Whether that demand will be able to be satisfied depends on whether doctors would be willing to supply larger volumes, but for this part of the population there would be a volume-increasing patient incentive. To the extent that Medigap policies cover the coinsurance, this effect will be attenuated.

For beneficiaries who use physicians who do not accept assignment, the potential implications are quite different. Suppose actual expenditures exceed projected expenditures, in which case the level of reimbursement per unit is cut with fixed total prices. The amount charged in excess of the Medicare-paid amount will then increase, compared to a non-expenditure-limit

program with the same nominal unit prices, and these charges (which would be in excess of a CPR maximum) will usually not be covered by Medigap. Higher user prices, in turn, would be expected to reduce the quantity of services patients demand. If physicians are profit maximizing (see below), this decrease in desired quantity would translate into actual volume decreases. Thus in this case, permitting balance billing (for those beneficiaries who choose to patronize such doctors) will add a patient incentive to reduce volume when overruns occur. Thus, for physicians who do not accept assignment the effort of a revenue limit is reinforced; this will not happen for participating doctors or those who frequently accept assignment. If physicians adopt a modified target income behavior, they might respond to lower prices by making further efforts to generate demand, but now demand generation will have to climb over both patients' natural reluctance and the higher user price they will be asked to pay.

In the target income case, the incentive effects on patients from a volume overrun will, for patients not billed on assignment, set in place a self-correcting mechanism. For patients of physicians who accept assignment or who are participating, the demand incentive is in the opposite direction and (in contrast to the predictions about physicians incentives) unequivocally so. As compared to a situation in which all physicians are compelled to accept lower prices (no balance billing), but patients pay 20 percent coinsurance, the effect of permitting balance billing is to offer a stronger patient incentive to correct volume overruns.

We suspect, however, that the incentive effects of balance billing will not be the most important policy question about balance billing. Instead, the fact that the cost of overruns, at least in the first instance, is shifted forward to patients will be regarded by many as objectionable. To some extent, the resolution of this question raises the same issues as are raised by questions of balance billing generally. Does one imagine that the people who are charged extra are people who could have used a doctor who accepts assignment, but decided instead to use one who does not in order to get some benefit whose cost they were quite willing to pay? That benefit might be more convenient access, higher amenity, higher technical quality, greater patient rapport, or snob appeal. Or does one imagine that the people being asked to pay more are unlucky and helpless individuals who have no other options?

We do not suppose that these questions can be answered in this specific context, any more than they can be answered in the more general one. There is some merit, one can argue, in permitting people to override Medicare's judgment about how much expenditures can rise, and how much technology can grow--as long as they know what they are doing and have other feasible alternatives. That is, it is in principle desirable to make it possible, even easy, for people to supplement privately whatever Medicare has decided to pay.

But one wants to limit the chances of an unlucky or uninformed consumer being gouged. The strategy here is probably no different from that in the general case. The program should make sure that there are sufficient acceptable alternatives available to beneficiaries in which additional amounts are not charged, and should also provide strong advice to beneficiaries about the availability of those alternatives. Some of the worst excesses could

probably be forbidden. But especially with expenditure caps, where mistakes by Medicare in forecasting are possible, and where people may have quite different preferences about what rate of growth they will tolerate, it is probably undesirable to outlaw the safety valve.

What can be done is to spotlight and isolate the high-priced providers. At a minimum, physicians not participating or accepting assignment could be subject to a special expenditure target, just for their services. Greater effort could be made to make beneficiaries aware that these doctors have not agreed with Medicare's judgment about the appropriate rate of growth in expenditures, and therefore, should be patronized only with full knowledge that they are different.

5.5.3. Summary

In the face of recent reductions and freezes in prices paid by Medicare for physician services, some have concluded that an explicit process of defining the target for Medicare expenditure growth would provide more predictability and control for both Medicare and the physician community. Expenditure targets for future time periods would be derived by determining the appropriate growth in aggregate expenditures for a given geographical region over a given period of time. Thus, expenditure targets would be defined based upon a population of patients, not a population of physicians. If expenditures were to exceed this target, then various mechanisms to change Medicare fee levels could be set in place to set total Medicare expenditures right, either in the current or in the next time period. Geographic regions could be established based upon the uniformity of input prices, variation in medical practice, the degree to which the physician community is likely to be able to exert influence, and a number of other characteristics.

All these models would predict that an expenditure target would enable Medicare to assure beneficiaries that their premiums and the outlays on behalf of their medical care would increase in a predictable fashion (presuming limits on balance billing). Expenditure growth targets cannot fail to control the growth in expenditures. They also would enable the medical profession, in keeping with the Patient Agency Model, to take responsibility for the control of volume and services and would forge new professional relationships and possibly new physician organizations. For example, with this volume control, peer interaction and review or case management programs, such as exist in many prepaid practices, might be instituted.

It would be desirable for the prices paid for medical services to be considered satisfactory by physicians before an expenditure target system was implemented, in order that physicians be willing or even eager to participate in such a system. Because there are, at present, perceptions of inequity in reimbursement among physician specialties (many of which would be addressed with the RBRVS), a system that makes one physician financially liable for the practice patterns of other physicians would likely meet significant resistance.

While expenditure targets could be mandatory, they could also include certain elements of voluntarism. Within a mandatory expenditure target system, groups of physicians could elect to "opt out" and to practice within subsystems of the expenditure target population. For example, health maintenance organizations or preferred provider organizations that have demonstrated their ability to provide care efficiently might choose to be providers with targets for their present populations in order to free them from the potentially higher utilization rates of the general physician and patient populations.

A number of design issues exist with regard to expenditure targets, including the method of projecting expenditures, the method of sharing overrun costs or savings, the timing of adjustments, the size of the geographic area, the degree to which assignment would need to be mandatory, and the issue of mandatory vs. voluntary participation. A number of concerns also exist, including the perceived fairness for physicians who are held accountable financially for the decisions of their peers, the potential that individual physicians would not limit their own utilization of services and that a spiral of increase in volume and decrease in price would occur.

Appendix: Theoretical Effects of Expenditure
Targets Under Behavioral Models

Appendix: Theoretical Effects of Expenditure Targets Under Behavioral Models

Expenditure Targets

Profit Maximization

In this model, physicians are paid on a fee-for-service basis. Total expenditure for a specified period and patient population are predetermined by the payer, and a nominal price calculated to induce a desired volume and expenditure is set.

We assume initially perfect information and predictability. Under this assumption expenditure target and price are correctly calculated to meet prevailing volume and physicians are able to forecast the volume and predict whether target overruns will occur. If the physicians accept the calculation and believe that no target overruns will occur, then profit maximizing physicians will produce the same volume produced prior to the implementation of expenditure targets. Expenditures will be fixed, and quality and access will not change.

It is possible, however, that physicians forecast cost overruns. In such a case, the forecasted price will be lower. The quantity supplied will therefore fall. With profit maximization, the impact of expenditure growth targets that are associated with overruns is unequivocally to reduce volume.

When target overruns occur next year, nominal price will be reduced and the paid price will also be reduced. This process will continue until the nominal price equals the marginal costs at a point when the target is exhausted. The system will reach an equilibrium. The expectation then will be that no budget overruns will occur. Expenditures will be fixed. However, excess demand may result. Reduced access to care in the profit maximizing model may or may not be a problem. Maximum induced-demand may reflect overutilization. Reduced volume may therefore improve quality of care.

In a two-sided expenditure target system it should be noted that physicians as a group could improve their welfare by producing the amount needed to maintain the original nominal price P_1 . They will work less and share the same amount of revenue. To obtain this result, collective action among physicians is necessary. The likelihood of such collective action will increase the smaller the physician group among which the targeted expenditure is divided. Thus targeting design considerations may indicate the advantages of expenditure targets aimed at small and specific groups (see section 5.2.2.). Theoretically, at least, this strategy can create access problems which may adversely affect quality of care. This will happen if the collective optimal volume of care is smaller than the true demand.

What happens in the "next period" model when there is an expectation of cost overruns in the current period? Next year's budget will be reduced by the amount of the expenditure overrun. Assuming perfect knowledge and full foresight the new expenditure target and nominal price (NP) will intersect with the supply curve at the point where $MC = NP$ and the target will be met.

Volume will decline and the expenditure target will be met (the assumption in this simplified analysis).

The discussion up to this point was based on the presumption that the purpose of the target is to lower the rate of growth but not to actually reduce expenditure. Expenditure targets may, however, be used to decrease expenditures by reducing induced demand, if it exists. If planners do not know the true demand function and the supply function they will drive the target expenditure and prices lower until a surplus is created indicating excess demand. In such a process physicians will have an incentive to collectively limit volume at the high target level and high nominal price. It is likely that physicians have better knowledge about costs and demand than do the planners. If collective action is organized it is important that the limits on volume will be in accordance with proper access and quality of care.

One-sided or Two-sided Expenditure Targets

The argument for a two-sided expenditure target is to provide incentives to create surpluses that will be shared among physicians and thereby reduce total expenditures and volume. In markets with high levels of induced demand such a policy may result in improved quality of care.

There is, however, a theoretical risk associated with this strategy. It creates an incentive to create surpluses by reducing volume. This strategy requires collective action. It is not clear that physicians could successfully organize and agree on an allocation system of volume and revenues. However, if successful, collective action may result in limited access to care and possible negative impact on quality of care. Two-sided systems should therefore be implemented in conjunction with a monitoring system reviewing access to care.

Conclusions

Expenditure targets that are correctly calculated will successfully attain the expenditure goals the system would like to achieve. However, since the determination of a target and a price result in a given quantity produced, it is important to know what is the quantity needed. One does not like to maintain a situation of overutilization with negative impacts on quality of care. Neither would one want to create excess demand which results in underutilization, restricted access to care and possibly reduced quality. Thus, in the long run expenditure targets should probably be implemented in conjunction with a quality monitoring system that adjusts pricing (if not expenditure) towards the provision of the right quantity. Government can help attain this goal by helping physicians to coordinate the allocation of revenues and volume based on proper medical criteria. Having a monitoring system to identify underutilization takes added importance in a double sided expenditure target system.

Expenditure Targets Under Sophisticated Income Targeting

Perfect Information and Predictability

An income targeting physician will not have a reason to change his or her decisions regarding volume and quality of care as long as the determined target-price combination guarantees his or her ability to maintain net income. In this decision we assume that the expenditure target makes provisions for justified and reasonable increases in net income due to inflation, population growth, and reasonable improvements in quality of life. Assuming full planner information and perfect physician foresight, the appropriate target expenditure, based on last year's expenditure level, could be calculated. Quality and access will not change. The growth of volume and expenditures will be successfully controlled.

Assume now that the income targeting physician expects expenditure overruns because of imperfect predictability. Unless physicians change behavior, they will suffer a decrease in net income. Under the withhold system the revenues from the withheld amount will be reduced. Under the "next period" system expenditure targets will be reduced as will price, revenues, volume and net income will decrease.

Given the set nominal price and/or expenditures, the income targeting physician may attempt to obtain the target income by reducing costs of inputs, thereby increasing net income. In addition, he or she may choose to create more demand and increase volume. The physician may choose to use either one of these strategies or both. Reducing the costs of inputs may adversely affect the quality of care. Efforts to induce demand will increase volume resulting in decreases in nominal price in the withhold model and decrease in expenditures next year in the next period model. Both adjustment will imply deviation from his previously preferred subjective method of practicing medicine. However, the model's assumptions allows for such a shift when net income decreases. The extent of the change and the type of adjustment likely to occur depends on the nature of technological substitutions available to the physician between inducing demand and reducing costs of service provision. The physician would choose the adjustment that reduces quality the least for a given increase in net income. He or she will also need to equalize the marginal benefit from net income and the subjective marginal satisfaction from professional practice style.

When expenditure overruns are expected and predictability is imperfect each physician has some incentive to ensure that net income is maximized before expenditure overruns occur and adjustments are made. In the withhold model as long as the paid price (the price prior to distribution of the withheld amount) is higher than MC, it is in the physicians financial interest to provide all the available demand. If he does not increase volume, perhaps other physicians, anticipating that the withheld amount will decrease, will satisfy the existing demand but he or she is constrained by the subjective cost. Consequently, the physician will lose income this year. In the next period if nominal price is not changed the physician would again forgo a portion of the withheld amount. If nominal price is reduced revenues will decline and so will net income. The same logic applies to the "next period"

system. It is in the interest of each physician to extract somewhat more income in the current year, if he or she expects next year's budget to be lower.

When reductions in net income are expected, both quality of care and/or volume may be affected. Whether volume is increased or quality reduced will depend on the subjective judgments made by the physicians. The imposition of an expenditure target may therefore induce overutilization in the short run. Services that were considered effective and beneficial prior to expenditure target might also become ineffective in the post expenditure target period. It is important therefore to implement the expenditure target policy with an effective monitoring system that monitors access and quality, if such a system is feasible.

Physicians as a group do not want the expenditure to shrink. Reduction implies less available revenues. In this environment, the incentives for collusion are much stronger than the incentives in the profit maximizing model because there is room to induce more demand. The more revenues available to the group, the better off they are. They can either induce demand or provide better care and derive more subjective professional satisfaction.

In the withhold model, if it is a two-sided model they would have an incentive to engage in collective action to limit volume and provide the minimal volume of services which uses all the available targeted expenditures as long as there is no excess demand. This strategy reduces volume. But physicians can obtain the same income while providing higher quality care to the patients they see. The reduction in volume in markets with high induced demand will improve care if physicians ensure that unnecessary procedures are avoided. Since their target income is achieved, there is no reason to believe that quality will not improve, given the assumptions of this model.

Under a two-sided expenditure target policy it is in the collective interest of the physician group to limit access to care and provide high quality care to the patients receiving care. This strategy provides the physicians with their target income, and with high level of subjective professional satisfaction. In addition, excess demand creates pressure to increase the amount of the expenditure target to provide care for those patients who cannot currently obtain care. To obtain this effect physicians will have to engage in collective action to limit volume and decide how to allocate the volume among themselves. This policy can be pursued under a two-sided system only because it could produce surpluses which physicians would like to share.

Expenditure Targets Under Patient Agency

While expenditure caps would be expected, like capitation, to have a profound impact on the organization of practice and on relationships among physicians, their direct impact on volume under this model of behavior would be minimal. If physicians as agents already provide what they view as optimal care without regard to their own income, the penalties associated with budget overruns (withholding of physician income in period 1 or a decrease in unit

prices in period 2--see Section D on Expenditure Caps) would have little, if any, influence on the care they deliver. Further, because neither penalty will affect beneficiary expenses, the economic agency aspect of physician behavior will not be challenged. In an extreme case in which the cap was constructed in such a way that once the budget limit was reached, no more services would be reimbursed in that period, a short-term increase in volume might be seen as each physician attempts to provide all needed services to his own patients early in the period.

Expenditure caps might influence volume if "low-utilizer" physicians exert pressure on "high-utilizer" physicians (where each believes his or her style of practice is best); in this case, the extent of the volume reduction would be consistent with the mix of high- and low-utilizers in the physician population.

5.6. Collapsed Coding

5.6.1. Overview of the Literature

To our knowledge, there is no empirical data on the use or effect of collapsed procedure codes on volume of services or expenditures, other than the limited findings reported in Section 6.0 of this report. Analysis of simulations of collapsed coding, using Medicare data from South Carolina are reported by Mitchell et al (1987).

5.6.2. Likely Effects of Implementation

There are thousands of different procedural codes that doctors can use to bill Medicare for Part B services. There is concern that this multiplicity of labels itself may contribute to high and growing Part B expenditures, and that reducing the number of billing codes "could control both costs and utilization by reducing the number of service units billed..." (OTA Report, p. 155.) In this section, we consider the use of collapsed procedure codes for similar services as a way of dealing with the problem of volume and intensity growth. In the next section, we examine the more far-reaching strategy of bundling services which are themselves quite different but which are usually used in combination for a visit, procedure, or episode of care.

Collapsed procedure codes (CPCs) represent a strategy of combining codes for similar procedures ("procedures that have only fine distinctions") into a single code, and paying the same reimbursement regardless of which procedure was actually used. Such a strategy might eventually reduce Medicare's administrative costs, although the transitional costs would be large, and the present value of net administrative cost savings (depending on the discount rate used) could be zero. It is even possible that administrative costs could rise, especially if (as discussed later) collapsed procedure coding is accompanied by more vigorous efforts to detect and deter inappropriately high coding. However, the major objective of this strategy is not to save on administrative expense, but to reduce the average costliness and/or number of services billed.

5.6.2.1. What Difference Do Labels Make?

The doctor knows with certainty what services were provided, and how useful those services would be for the patient's problem. The patient observes some (though not necessarily all) of the services, and may be able to form an opinion about the types of services actually rendered. (We consider below whether and how beneficiaries might help to prevent upcoding.) HCFA or its agent observes a bill that is submitted, with no knowledge of what services actually were rendered, what the patient's condition was, or what the outcome was. This means that, regardless of the coding scheme, HCFA's ability to detect manipulation of procedure codes will be limited. One strategy would be a system of spot checks to detect and punish specific cases of incorrect coding. From HCFA's point of view, the deterrent effect of any such specific enforcement policy will depend on the punishment it can impose.

The alternative method of deterring incorrect coding is by monitoring the relative levels and changes in the frequency of high-revenue codes, and imposing general penalties for unusual changes, without regard to proof of specific error. We will not consider this strategy in detail here, other than to note that (a) it is probably sensible to monitor even if all coding is accurate and (b) the method can be rendered ineffective by gradual and parallel changes in coding.

Medicare's objective with regard to coding is to pay only for those services rendered, and to pay the price which it has already determined to be appropriate for that service. The problem is clear: there is no way for HCFA, or the patient, to actually determine what services were rendered or whether they were appropriate; there is always a possibility of some lack of correspondence between what is billed and the appropriate payment for a particular service.

Collapsed procedure codes represent an attempt to deal with one part of this dilemma. We will define CPCs as directed at situations in which the doctor must choose to bill for one service from a set of similar services. In contrast, the "bundling" issue deals with the choice of the mix of a set of dissimilar services which are or can be used in combination to treat a particular patient for a particular condition.

There are then two fundamental choices the doctor must make in deciding how to deal with a Medicare patient who would be a suitable candidate for one service from a set of similar services. The doctor must decide:

1. what to do for the patient; and,
2. which service to bill for.

Problems arise because the actions the doctor takes do not always have an obvious and automatic match to one of the procedure codes under which he or she must bill. Problems also arise, even when the assignment of codes to services is obvious, because the doctor must decide which service to render. That is, there are problems of both labelling and choice. In terms of HCFA's expenditures, these translate into problems of procedure inflation and (true) high-cost volume expansion. The thought of those who favor CPCs is that altering the set of possible procedure codes may solve one or both of these problems.

A note here on terminology. In this report we will limit the use of the term "procedure inflation," as originally suggested by Mitchell et al., to situations in which the same service is billed under a different and more remunerative code. This contrasts with the use of the term "upcoding" in the recent Mitchell et al., report on the impact of PPS, in which any shift in the mix of set of similar services toward billing under more lucrative codes was called "upcoding," without regard to evidence on whether the content of the services had changed. For instance, a shift in the mix of colonoscopy bills toward procedures involving endoscopy beyond the splenic flexure was called upcoding, but there was no evidence offered that in fact this shift misrepresented what was actually done. "Upcoding" therefore seems to refer to both procedure inflation and high-cost volume expansion.

5.6.2.2. Physician Behavior and Procedure Codes

It is obvious that answering the question of the possible impact and desirability of CPCs requires answering a question which has as yet not been addressed: one needs to specify how one thinks a doctor decides on the billing procedure code to be used for a given service. The method used by the doctor's office staff to arrive at a billing code also needs to be specified. We therefore begin with the behavioral question, and we do so in the context of a single physician who has the power to direct how his or her services will be billed.

We think it is important first to distinguish between technical procedures and visits. It seems reasonable to suppose that the actual performance of a technical procedure could, in large part, be known to the patient, and could, in principle, be monitored by HCFA. If a physician bills for a fiberoptic colonoscopy beyond the splenic flexure to remove polyps, he either did or did not perform this task. Moreover, if there was a polypectomy, tissue will have been removed. In contrast, if he or she bills for an intermediate office visit, there is no specific set of tasks to be performed; the service is defined primarily by the type of patient seen and the type of problem presented. While there has been some thought given to redefining visits in terms of specific sets of procedures, at present, it would be very difficult to judge whether the "service" really was provided--beyond noting that a visit has occurred.

We begin by considering the office visit case. The physician sees a patient for thirty minutes; he has seen this patient before in the distant past for a condition possibly somewhat related to the symptoms the patient now reports. How shall he bill the visit--limited or intermediate? The one clear message is that there is a great deal of uncertainty about which label is correct. If the physician formerly billed for such visits as "limited," and now switches to "intermediate," we surely could not call that fraud; neither external observers nor the physician could resolve the fundamental ambiguity about what the label should be. Instead, the reasonable question to ask is what determines what label the physician will attach. We will return to the "visit" case specifically below.

In contrast, in the case of technical procedures there is much less intrinsic ambiguity (though there may still be some). The doctor may in some circumstances nevertheless choose to mislabel, either for his or her own benefit or for that of the patient. It is important to note that procedure inflation which increases the level of Medicare payment may benefit the patient; the upcoding physician may be more willing to accept assignment. The critical question here is the (negative) value attached to such fraudulent billing.

There are two senses in which procedure inflation might be limited. In much the same way as for demand inducement, one might suppose that the doctor suffers a subjective cost from stretching definitions in order to bill at a more lucrative level. He or she may be willing to move to an adjacent procedure code for which classification might, in any case, be debatable, but he or she would be less willing to make so large a movement that it could

clearly be labelled fraud. Alternatively, the possibility that procedure inflation would be detected, and penalties imposed, may also act as a deterrent. The detection could either come from the patient or from a Medicare's own auditing activities. The deterrent obviously depends both on the probability of being detected and the penalty that is imposed if procedure inflation is detected.

How would a change in the number of technical procedure categories be expected to affect the extent of procedure inflation? By "extent" here we will mean the difference, measured in dollars, between actual billing and the billing that would have occurred had the services been correctly reported. Mitchell et al., (1988) argue that "with a greater number of procedures to choose from, the physician has more latitude in billing under a more complex, costly procedure code for the same service." This statement is obviously true if the alternative is just one code, since then there is no choice. If the number of codes is large, the notion appears to be that "it is easier to subjectively rename" simple procedures as slightly more complex ones when a "close" code exists. The intuition is apparently one that attaches subjective (disutility) cost to procedure inflation, with the size of the cost depending on some notion of the distance from what is reported to what was done.

However, there is more to coding behavior than just the "ease of renaming". Except for reducing the number of codes to one, any other reduction in the number of codes will not necessarily reduce the extent of procedure inflation. Procedure inflation (measured in dollars) could be reduced if the number of codes were cut, but it could also increase. The ambiguity arises for two reasons. First, we do not know how (or whether) the willingness of the doctor to bill fraudulently varies with the "space" between codes. Second, even if upgrading becomes less likely when codes are condensed in a budget neutral situation, the average dollar value of an upgrade will increase for those who do upgrade. If the number of codes is halved, the frequency of upgrading might fall by half--but if the dollar cost of an upgrade doubles, the net amount of procedure inflation, measured in dollars of Medicare expenditures, remains exactly the same.

What determines the frequency of upgrading? A larger step between categories presumably makes it more likely that HCFA may notice--and punish the doctor. It also makes it more likely that the patient will notice, and at least raise concern. Finally, it is likely that the subjective cost of modifying codes will increase with the size of the step. Although the subjective cost to the doctor of upgrading therefore increases, the reward for doing so also increases if reduced coding is budget neutral. Hence, it does not seem possible even to conclude that there will be less upgrading. Reducing the procedure inflation cost requires not only that the frequency of upgrading fall, but also that it fall by enough to offset the increase in the cost for upgrades which do occur.

The real issue is the rate at which physicians are willing to upgrade per dollar of financial reward for doing so. If the extra revenue for upcoding is less than the subjective cost of doing so under multiple codes, then reducing the number of codes can actually lead to more procedure inflation. Reducing

the number of codes by half will, on average, double the value of upgrading, and may well induce more such behavior.

The appendix to this chapter develops some numerical examples to illustrate this point. We show there that "continuous coding" may in some circumstances actually be the strategy that eliminates procedure inflation.

We do not want to push this argument too far. If codes are collapsed so that upcoding will be obvious to everyone, and penalties easy to impose, then collapsed coding will deter procedure inflation. In a sense, the "technology" of detection and punishment determines how well collapsed coding will solve the problem of procedure inflation. The answer also depends on the true distribution of procedures. If most procedures are in reality clustered about a "medium" procedure, with relatively few truly complex or very simple procedures, then defining three codes--say, "simple, average and complex" will reduce measured volume and intensity. Conversely, if services are distributed near the two extremes, collapsed coding may lead to higher spending.

It might, however, be possible to encourage beneficiaries to detect and report procedure inflation. Private insurers, such as Aetna, have, in fact, instituted programs urging patients to communicate to the company information about bills for services which differ from those provided, on the basis that such coding is detrimental to insureds in the form of higher premiums.

Such education and encouragement might also help in the Medicare program, though the complex nature of Medicare billing could generate many "false positives" that will be costly to investigate. Substantial effects will, however, probably require stronger incentives to beneficiaries. One could imagine a process of offering financial rewards (reduced cost sharing) for reported overbilling. Many patients will, one suspects, be reluctant to report on their doctor, both because of the doctor-patient relationship and because overbilling will, other things equal, probably make the doctor more willing to supply the services he or she does provide, and to accept assignment. Some investigation of a program to encourage beneficiaries to check their bills more carefully may be warranted.

5.6.2.3. Specific Cases

The "bimodal" distribution of procedures described above may well characterize the distribution of ambulatory care visits. There has been a pattern of upcoding of ambulatory visits--although the question of whether that represents a change in billing or a change in visit content makes it difficult to evaluate. In either case, a useful modification might be one which reduced the number of visit codes, and eliminated differences in price depending on where the visit took place. Specific documentation of need and content would be required for more complex visits.

As an alternative, the price paid could simply depend on the length of the visit, something which could probably be monitored more easily than content. Paying per minute could be more compatible with RBRVS.

In contrast, reducing the number of codes for colonoscopy or gall bladder surgery might be expected to lead to more glaring attempts to upcode. Rather than reduce the number of codes, it will be more important to get the price right for each procedure.

5.6.2.4. CPCs and High-Cost Volume Increases

How would moving to CPCs affect true volumes of care? Would reducing the number of codes lead to a reversal of volume expansion? The answer depends in the first instance on whether the doctor is assumed to have appreciable ability to create demand. If this ability is limited, then there is nothing to be gained or lost from changing codes, unless there is excess demand for some services. If demand is satisfied, the quantity would not be expected to change.

But suppose the doctor has some ability to alter the volume of services received. Moving to CPCs in a balanced budget way reduces price for the high cost (formerly high-priced) services and increases the price for the low cost (formerly low-priced) services. The "substitution" incentive would indeed predict that doctors would want to create less demand for the formerly high-priced services or even ration existing demand. But a symmetrical argument necessarily implies that they would want to create more demand for the formerly low-priced services. What will happen depends on how hard it is to destroy demand, and how easy it is to create demand for each type of service; something we know very little about. Since the lower-priced services typically use less patient time, one might conjecture that demand creation would be easier there, so that movement to a CPC system would actually increase aggregate dollar volume. But given our present state of empirical knowledge, little more can be said.

In fact, the question about CPCs is equivalent to asking the more general question of what happens when one changes the relative values for a related batch of medical services which are close substitutes for each other; it is the same question as asking about the impact on volume of the resource-based relative value scale. After all, paying the same price for different services which use different amounts of physician time is contradictory to the policy of basing price on relative physician time input. As we show in another section of this report, such a change has no predictable impact on total billed volume; it is as consistent with an increase as with a decrease.

Mitchell et al., investigates a CPC system which significantly changed average Medicare prices received across specialties. The significant reduction in income a specialty might experience could lead to an increase in volume of the services it provides.

5.6.2.5. CPCs and Patient Demand

Introduction of CPCs will change the copayments beneficiaries pay and, if assignment is not accepted, the amount billed to the patient in excess of the Medicare payment. Copayment will fall for the formerly high priced services,

and rise for the formerly low priced services. These changes in copayment will affect beneficiaries who do not have Medigap coverage. One would therefore expect the quantity demanded of low priced services to fall and that for high priced services to rise, exactly the opposite of the expected incentive effects on doctors. However, if doctors are more likely to refuse assignment on high cost services, and accept assignment on low cost services, the effect of balance billing will offset the effect of copayment. That is, the high cost service whose Medicare reimbursement level is cut will experience an increase in balance billing, which may offset the lower copayment on the portion which Medicare pays.

The effect on patient demand of any change in procedure inflation is less clear. If the net effect of having only a few categories is to reduce procedure inflation, and therefore the prices doctors receive, there will also be an offsetting copayment effect. This offsetting effect will stimulate patient demand, and increase patient willingness to accept larger volumes of service.

5.6.2.6. Formulating Policy in a World of Uncertainty

Without empirical evidence on the effect of the number of billing codes on expenditures, there is no reason to expect CPCs to slow volume growth. Reducing the number of codes could just as well increase volume; we just do not know. More generally, if one could monitor to make sure that what is billed is what was provided, it would seem that more, not less, procedure refinement would be appropriate. After all, if two services are different in the sense of having differing costs, ideal pricing would require that they receive different prices. While RBRVS does not necessarily move toward ideal pricing, its logic would nevertheless seem to require that procedures with different costs likewise be treated differently.

In the meantime, what is needed here is some more direct study of the empirical effect of number of codes on reported volume. The variety of systems used by public and private carriers and the frequency of changes in coding should provide a rich set of natural experiments that would answer empirically the question that is theoretically ambiguous. An alternative to collapsed procedures codes would be to direct efforts at monitoring services actually provided, given current coding, and punishing severely any fraudulent billing. Where there is honest ambiguity about the definition of services (as in the case of visits), it might be well to recognize that trying to measure the unmeasurable is likely to do more harm than good. Paying doctors per visit on a per minute basis, and then monitoring to make sure that there is no excessive visits or excessively long visits might be preferable. Here again, theory is likely to be inconclusive; evidence will be needed.

It would also be desirable to pay careful attention to the introduction of new procedure codes. Permitting a new code should be based, among other things, on a proven ability to monitor to prevent procedure inflation with that new code. One should also determine whether the new code represents different levels of resource cost than the current alternatives. If cost is

the same, defining a new code only generates more opportunity for confusion. Clinical differences between services will also be important.

Collapsing procedure codes definitely leads to warped incentives for doctors who are telling the truth, and has a problematic impact on expenditures for those who are willing to engage in procedure inflation. There would be a saving in administrative cost, and a reduction in confusion, but an increase in frustration by doctors. On balance, however, it does not appear that CPCs per se can have a major impact on expenditure growth.

5.6.3. Summary

Given the large number of codes for medical services, it is possible that physicians may choose to "upcode" by indicating that the service they provided was one with a slightly more complicated level of severity or intensity. Thus, collapsing procedures into fewer codes could potentially reduce the opportunity for upcoding. However, our analysis suggests that the net effect of collapsing codes is unpredictable. In essence, collapsing codes in balanced-budget fashion reduces the incentive to upcode by increasing the "distance" between two codes, but this may be offset by the increased incentive to "jump the gap" due to the higher price differential between two codes.

Appendix: Theoretical Effects of Collapsed
Procedure Codes Under Behavioral Models

Appendix: Theoretical Effects of Collapsed Procedure Codes Under Behavioral Models

Collapsed Procedure Codes Under Profit Maximization

A. Technical Services - Clear Definition

In this volume control instrument it is important to distinguish between profit maximizing behavior and opportunistic behavior. Profit maximizing behavior is aimed at maximizing the difference between revenues and costs. This behavior does not include fraud. Opportunistic behavior is morally questionable behavior which exploits opportunities for personal gain and while it may be applied towards gaining extra income, it is not restricted to increasing net income.

Assume that a service A has S substitutes $A_1 \dots A_5$ for which the reimbursement rates increase by \$10 as we move from A_1 to A_5 . The costs increase by a lower rate. We also assume that A_5 includes all the properties provided by A_4 and some additional services. The profit maximizing physician will provide the service which maximizes net income A_5 . That is, we are likely to observe high cost volume expansion but no procedure inflation.

Collapsing procedures and paying an average fee will reduce the volume of A_5 since its profitability will decrease but it will increase the profitability of A_1 whose price was below the average price. The volume of A_1 will increase. The volume of A_5 will decline and total expenditure will decline. Total volume will remain the same, but the procedure for which the reimbursement rate is above average, and whose reimbursement rate declined will not be provided, if they are less profitable. Total volume will remain the same because maximum demand inducement will occur before and after collapsing.

In this model, if physicians were observed to provide services other than A_5 it is because they were not able to create demand for A_5 . If after collapsing a service falls into a category for which its price is above the reimbursement rate, procedure inflation could not occur, because it was impossible to induce demand for the next level. This service will therefore not be offered. The services whose old price was below or equal to the new reimbursement rate will be offered at the new price.

While expenditures will be reduced, overall quality of care could improve. This will occur for those services where prior to collapsing overutilization of the high reimbursement items prevailed and resulted in negative impact on quality. Since these services will no longer be provided the negative impact will be reduced. However, some patients who need these services will not be able to obtain them, resulting in diminished access and a decline in quality.

B. Non-technical Services - Unclear Definition

As explained in the previous section under this model, high cost volume expansion will occur resulting in provision of the highest reimbursed service at the highest possible volume. Procedure inflation will therefore not be a problem unless demand creation is impossible. If demand creation is impossible procedure inflation will not occur. The service provided will always be the most profitable given the ability to create demand. Thus if A_3 was provided for \$30 it was because A_4 for \$40 could not be provided. If under collapsing A_4 is now reimbursed at \$45 as is A_5 it would be impossible to charge \$45 for A_3 (procedure inflation). Consequently, A_3 will not be offered. A_2 will be provided at, say, the reimbursement level \$20.

In this model, collapsing codes will result in lower expenditure. The impact on quality of care is not clear since we do not know what is provided and what difference in quality is obtained by purchasing A_2 vs A_3 , for example. Access for the type of service will not be affected.

If enforcement and penalties are so strict as to deter procedure inflation physicians will have to choose between high cost volume expansion or eliminating access to services whose prior reimbursement rates were above the rate currently reimbursed for their category. This will result in reduced quality and, limited access to care. Total expenditure cannot be predicted.

Collapsed Procedures Under Income Targeting

A. Technical Services - Clear Definition

In this model, we expect to observe a wider distribution of services provided, compared to the profit maximization model. Physicians will be more reluctant to induce demand for the high profit services if it is not consistent with quality care. Collapsing codes would not change the physicians medical choice. It would affect his or her coding practices.

Assume A_3 was provided at \$30 per unit and is now collapsed to a category reimbursed at \$20. A_3 will be provided coded as A_4 and reimbursed at, say, \$45. Physicians could do this because they can induce demand for A_4 . Inducing demand may be more costly and impose some subjective costs regarding honest reporting. However, the appropriate clinical procedure was provided. The extra charge from Medicare can be balanced with decline to except the coinsurance of deductible portion from the patient. Physicians who did not accept assignment may be ready to accept assignment while practicing procedure-inflation.

In this model, less high cost volume expansion will occur, but more procedure inflation will occur. Since the physician has a commitment to the provision of proper care, and since he or she can provide what is considered proper care, if he or she upcodes those services that are deemed necessary, upcoding would be the less of two evils. Assuming no more opportunistic behavior than before the collapsing option was implemented, we may observe an increase in expenditure due to collapsing. The services whose reimbursement

rates were below average will be provided and charged at a higher rate. The services with a price above the current reimbursement rate for their category will be charged at a higher rate (procedure inflation volume will remain the same and total expenditures will increase). Quality will be unchanged and so will access. Physicians may use the extra expenditures in two ways. They can pocket the excess or relieve the patients' burden of out-of-pocket expenses.

B. Non-technical Services - Unclear Definition

Collapsing procedures will not have impact on high cost volume expansion because clinical judgment will not change. Procedure inflation will occur because physicians will not be willing to forego income. They are likely to provide some care and charge more for the services whose reimbursement rates have declined.

In this case, the risk of penalties is lower than in case A. Thus procedure inflation is more likely. Total expenditures will rise, volume, access and quality will remain unchanged.

Appendix: The Relationship of the
Number of Procedure Codes to Total
Medicare Cost: A Numerical Example

Appendix: The Relationship of the Number of Procedure Codes to Total Medicare Cost: A Numerical Example

Introduction

We argued in the text that the relationship between the number of procedural codes and the dollar value of "procedure inflation" is ambiguous. Under a reasonable set of assumptions about how doctors decide what procedure code to assign, increasing the number of available codes can as easily reduce Medicare expenditures as increase them. We illustrate the possibility that expenditures can be reduced by more codes with a simple numerical example.

Assumptions

How much procedure inflation will occur obviously depends on the distribution of services by "true" codes (the code that is appropriate to the service actually delivered), the closeness of more lucrative codes for any service, and the willingness of the doctor to write down the more lucrative code rather than the true code. The doctor may be unwilling to write down a more lucrative code either because of the possibility of detection and punishment or because he or she feels guilty about stretching definitions too far.

We therefore assume for our simple example that:

- 1) Services rated by true codes are uniformly distributed by complexity "units" over an interval from 10 to 30. That is, there are 30 different "true" services.
- 2) Regardless of the number of codes permitted, the price of a complexity unit is \$1.
- 3) The doctor is willing to increase the stated complexity of a service (compared to the "true" value) by 0.5 unit for every \$1 in additional revenue.

The last assumption is a simple way of specifying a model of the doctor's willingness to upcode. It says in effect that the doctor requires \$2 in revenue to upcode by one unit, \$10 to upcode by 5 units, etc. The reason why the doctor would not upcode for any smaller gain could either be because he is ethically unwilling to do so, or because the probability of being detected and punished in some way rises at a rate that make the risk worth taking only for the stated amount of money. This relationship is specified as linear here; it might more reasonably display an increasing marginal "cost" of upcoding, a possibility we will discuss below.

Case A: Two Prices

Assume that two codes are defined, one at 15 units and one at 25 units, with prices of \$15 and \$25 respectively. Billing for any service is supposed

to be set at the closest available code. For instance , services of complexity between 10 and 20 are supposed to be billed at \$15, and those between 20 and 30 are supposed to be billed at \$25. If all doctors billed all services exactly correctly, average Medicare expenditures would be \$20.

Now suppose that doctors are willing to inflate procedures. Assumption (3) implies that services from 15 to 20 will be upcoded to 20. Upcoding a service from 15 to 20 gains an additional \$10, but requires moving the coding 5 units up from its true value. Since one quarter of all services fall in the range between 15 and 20, it follows that one quarter of all services are upcoded. Consequently, the average price Medicare pays rises to \$22.50.

Case B: Four Prices

Let the number of different acceptable codes be increased to four, set at 12.50, 17.50, 22.50, and 27.50. In this case, services with true codes from 10 to 15 are to be billed at \$12.50, those from 15 to 20 at \$17.50, those from 20 to 25 at \$22.50, and those from 25 to 30 at \$27.50. If procedure inflation occurs, services from 12.5 to 15 are upcoded to 15, for a gain of \$5.00. Likewise, services from 17.5 to 20 are upcoded to 20, and those from 22.5 to 25 are upcoded to 25.

So one half of the services in three quarters of the range of services are upcoded; that is, three eighths of the services are upcoded for an average gain of \$5. Hence, the average payment becomes \$21.875.

Comparing case A and Case B, we can see that an increase in the number of codes does indeed lead to more procedure inflation; upcoding becomes "easier." But because the gain from upcoding is less, it can happen, and indeed does happen here, that the amount of procedure inflation measured in dollars is less in the case with more codes than in the case with fewer codes. Medicare expenditures could be cut by adding more codes.

Case C: 30 codes

We now assume that there are 30 different codes, one each for the set of 30 services. Upcoding by one unit will only gain \$1.00, not enough to make upcoding worthwhile. Doing more upcoding will not help, since the marginal gain from upcoding is always less than the minimum gain required to cover the "cost" to the doctor of upcoding.

In this example we get a striking result: if doctors are willing to engage in procedure inflation, the best way to prevent that is to have as many codes as there are different possible procedures. The more codes, the better.

Conclusion

This simple example shows that more codes do not necessarily increase Medicare's expenditures. The answer comes from the assumption that the

"marginal cost" of procedure inflation by one unit is less than the marginal benefits. So procedure inflation only occurs when there are large jumps in revenue as a result of upcoding, and the size of such jumps is minimized by having many codes. Of course, if the marginal cost of upcoding starts out low and then increases, this conclusion may not hold. But the general proposition--that the determination of whether reducing the number of codes will save money is an empirical question, which cannot be settled by theory and does not have an obvious answer--still does hold.

5.7. Service Bundling

5.7.1. Overview of the Literature

To our knowledge, there is no empirical evidence on the use or effectiveness of service bundling in controlling volume and expenditures, other than the limited information provided in Section 6.0 of this report. Conceptual discussions of bundling can be found in Mitchell, et al, 1987; Mitchell, 1985; and Jencks and Dobson, 1985.

5.7.2. Likely Effects of Implementation

The increased volume and intensity in Part B services that has been experienced over the last few years has not been associated with an appreciable increase in the number of "contacts" or episodes of care for Medicare patients. The number of outpatient/ambulatory visits per beneficiary has barely changed. To a considerable extent, the growth in volume and intensity has been a reflection of growth in the amount and complexity of services per basic unit of encounter -- more lab and diagnostic tests per ambulatory visit. While it can be argued that the case mix of ambulatory patients has increased in severity recently, the question of whether at least some of these increased services represent physician-induced demand remains. Bundling of services would be expected to decrease the incentive to induce demand for complementary services.

In the current fee-for-service system, there is no defined "output" for physician services. The procedure codes used essentially identify inputs to the process of caring for a patient. There are no effective limits to the number of providers ("suppliers" of the inputs) that become involved in the process of care, or to the number or types of services each provider supplies. Medicare pays each provider for the input services supplied without regard to what health care "output" is being provided, or to whether all the services add up to an efficient provision of that service output. Total outlays for the Medicare Part B program depend largely on the number of providers involved in the process of care and the types and quantities of input services they provide. At current levels of prices, physicians probably have a positive incentive to expand volume and intensity.

The most comprehensive way of getting control over the cost associated with the total volume and intensity of services is, of course, per-person, per-time period capitation for the total set of Medicare services, inpatient and outpatient, physician and non-physician. Under capitation, payment is independent of the number of providers involved and the amount of service they supply. While Medicare has taken some steps in this direction, it must be recognized that, at least for a time, Americans are not ready for mandatory global capitation. The institutional structure needed to implement full capitation everywhere for everyone does not yet exist. On the supply side, a complex set of between-supplier transactions would be needed to divide a capitated payment in an efficient way. On the demand side, beneficiaries and policymakers rightly are concerned about the consequences of the strong financial incentives to contain cost embodied in capitation; it is easy to overshoot and go from reduced provision of unnecessary or marginal services to

the restriction of services whose benefits are worth their cost. Capitation therefore requires a structure for quality assurance.

As an intermediate step between full fee-for-service and full capitation, schemes to bundle services in units less inclusive than full capitation might be considered. These units must be more inclusive than fee-for-service with its thousands of different procedure codes. The bundling strategy is to identify some type of "main" unit that approximates or correlates with an output or intermediate good -- either an important procedure or an episode of illness--and bundle payment for that unit with that for the other services that commonly accompany it. These other services along with the "main" unit, can be viewed as inputs for the "output bundle." The various ways that this might be done can be seen as different ways of reducing the dependence of payments on the number of providers and/or the amount of input services they supply.

The major design problem is that there are many ways to cut up and rebundle the full set of services. In what follows we will propose a set of principles which might be used to make this choice, and illustrate the type of bundling that follows from application of those principles.

5.7.2.1. Objectives For a Bundling Strategy

The main objective for bundling is assumed to be control of the type or number of services provided. Incentives for choosing a lower cost (or price) provider of the same service are also, in theory, present in capitation or any other bundling approach, but this chapter will assume that control of differences in price across providers is accomplished by other means.

A major objective for physician payment mechanisms, as discussed earlier in this report, is "incentive neutrality," an arrangement in which the doctor is financially indifferent between alternative ways of treating a patient, in the sense that his or her real net income is unaffected by the action he chooses. The notion is that the physician will then be most likely to choose the action which provides maximum net benefit to the patient. While incentive neutrality has considerable appeal (implicit in the discussion of RBRVS), incentive neutrality is still consistent with the provision of very costly services of marginal but positive benefit to the patient, for which no lower cost substitutes exist. That is, under incentive neutrality the doctor will not be motivated to resist providing services which benefit patients modestly, but at great cost; he will not be induced to limit "moral hazard." In contrast, under full capitation there is an incentive to avoid providing costly services regardless of benefit. To the extent that capitation deviates from incentive neutrality in a negative direction (too few services), capitation's appeal is lessened even though it may be more effective than a neutral system in controlling cost. We therefore regard incentive neutrality as a useful benchmark.

Neutrality can be relative (as in an ideal RBRVS, before consideration of a multiplier), in which each physician "action" yields the same positive profit (but there is more profit from more actions in total). Or it can be

absolute, in which any action and "no action" leave the decisionmaking doctor with the same real net income. The ideal incentive structure is to get as close to absolute incentive neutrality as possible. As noted, pursuit of this objective need not lead to as low a level of Medicare expenditures as mechanisms which reduce real physician income when additional services are provided (have "negative net marginal income")--for instance, global capitation. However, we regard pursuit of the less powerful incentive structure represented by neutrality as appropriate, at least until decisions are made on how low a level of intensity is just low enough.

We imagine that any "bundled" physician payment system must define a major service. This main service could either be a specific procedure, such as a colonoscopy, or the batch of services rendered to treat a particular illness over a particular time period. A particular physician, called the "principal physician," provides this service and is to be the recipient of the bundled payment.

There are three categories of bundled services:

(1) Principal physician services. These services are almost always provided in fee-for-service practice by the principal physician if they are provided at all. Examples include follow-up visits and pre-procedure evaluations.

(2) Referred services. These services are almost always referred to a different physician or provider from the principal physician, if they are provided at all. Examples would be anesthesia or anatomic pathology services.

(3) Discretionary services. These services are "discretionary" in two senses: they may or may not be performed and, if they are performed, they can be rendered either by the principal physician or by a referral physician. Examples include lab tests and sigmoidoscopies.

There is a reason for distinguishing these three categories of services. As far as the principal physician is concerned, referral services are incentive neutral. That is, he or she neither gains nor loses financially from recommending these services. There can be financial incentives if there is fee splitting, if the referral physicians are both members of a multi-specialty group practice that divides income, if the physician is a shareholder in a service that he or she refers patients for, or if the referral leads to an increase in patient satisfaction that results in the patient returning for further care. But otherwise the physician is subject to incentive-neutral prices for such services whether they are bundled or not.

In contrast, if price exceeds real marginal opportunity cost, there is a positive incentive to the principal physician to supply principal physician services under conventional fee-for-service. He or she receives more gross revenue from doing so. If practice costs are not too great, and if the revenue from withdrawing his time from the non-Medicare market is not too high, profit will be increased by providing more such services. Especially under the target income motivation, the level of such services may therefore

be higher than under a bundling alternative with a total price that covers all major physician services.

5.7.2.2. Approximations and Incentives

It is the third category, discretionary services, that are the most problematic. If the principal physician is to receive a bundled payment, and if he or she then refers to other doctors for some of these services, he or she will necessarily engage in fee splitting, something which, at a minimum, will require a new set of business relationships among doctors, and something which, in many states, is technically a violation of the law.

The reason for the legal apprehension is concern for the incentives in bundling: decisions on whether and to whom to refer will depend on income to the physician (which is based on price and cost) as well as patient well-being; they will not be incentive-neutral.

Remaining with conventional fee-for-service payment does not, however, necessarily make matters better, since the incentives for discretionary services are not neutral under itemized fee-for-service either. The principal physician who is paid on a fee-for-service basis may choose to perform services himself or herself, as long as there is a positive net income, rather than refer to a lower cost and more adept referral partner. (Pauly, "The Ethics and Economics of Kickbacks and Fee-Splitting," Bell Journal of Economics, Spring, 1979.) There is, in short, a tradeoff. If discretionary services are included under bundling, the principal physician will have a positive financial incentive to refer to the lowest cost referral partner, if such services are to be performed at all, but to avoid providing such services. Under fee-for-service, the principal physician will want to provide the services, but may want to provide too many of them personally, even when he or she is not the best person to do so.

If bundling does not permit the principal physician to gain revenue from referred discretionary services, the physician will have an incentive to refer excessively; there is less of a chance of underprovision than under more inclusive bundling, and perhaps a smaller chance of overprovision than under fee-for-service. It is logically impossible to have both incentives to minimize cost and incentives to maximize patient receipt of services.

Some compromise must be sought. A sensible strategy would be one which divided the referral services into two groups based on evidence of whether, on average, they were performed by the principal physician or by other physicians. Payment for the former type of service would be included in the package, and separate billing would not be permitted. The latter services, in contrast, could continue to be permitted to be billed separately.

A somewhat different approach would be one based on the comparative cost of (quality-adjusted) services: when the principal physician's cost for some service is close to that of the referral partner, the service could be included in the package.

5.7.2.3. The Basis of Bundling

Is there any reason to choose one basis for combining services rather than another? Five principles should govern. First, administrative costs of arranging fee-division should be taken into account. Second, services provided by the principal physician definitely should be included in the bundle. Third, the more a service is substitutable between those performed by the principal physician and those performed by other input suppliers, the more appropriate it may be to include it in the bundle. Fourth, the less the variation in severity of illness that accompanies the major service, the more appropriate it will be to bundle. And fifth, the greater the prospects for demand creation, the more useful it will be to bundle.

What about a basis for excluding other related services from the bundle? Are there services that might not need to be included? If some service (other than the major service) provided by the principal physician is strongly complementary with other referred or recommended services, there is no need to include those complementary services in the bundle. Controlling the principal physician services alone will suffice.

For example, suppose a medical diagnostic procedure is usually followed by a number of physician visits. Each additional physician visit leads to more diagnostic tests, consultations, and drug prescriptions. One could bundle all these services together, but that would require that the principal physician divide a large payment, and could subject him or her to large losses in the case of a severely ill patient. In contrast, if only the follow-up doctor visits were bundled with the major procedure, the principal physician would not be at as large a risk, and would not need to split fees. And yet, such bundling would be fully as effective as the more inclusive but less feasible bundling in affecting the volume of complementary services.

This is admittedly an extreme case of complementarity. However, even less strong sources of complementarity can serve as the basis for a useful bundling exclusion. At one level, for physician services which have non-physician complements, there is less need to include the complementary services in the bundle.

An example may help to illustrate these principles. Suppose a primary service can be accompanied by:

- (1) Follow-up visits of variable number, always provided by the primary physician (own-provided service);
- (2) Lab tests provided by the principal physician, fixed in number in relationship to the follow-up visits (strong complements);
- (3) Other diagnostic services, which accompany the follow-up visits about half of the time (weak complements);
- (4) Consultations which are good substitutes for follow-up visits (discretionary services);

- (5) Home health visits, which are sometimes ordered by the principal physician but which he does not provide and which are not substitutes for his or her own services (referral services).

If incentive neutrality is the objective, the home health visits need not be included in the "bundle." (They might also be excluded on grounds of weak complementarity.) Because of strong complementarity, there is also no reason to include the lab tests in this example. In contrast, the own-provided follow-up visits definitely should be included in the bundle.

The more difficult cases are those services which are weak complements or which are discretionary services. The weak complements could probably also be excluded from the bundle under the rubric of incentive neutrality. If, however, it is felt that the volume of these services is sometimes excessive (for instance, suppose clinical evidence of benefit is present only one quarter of the time), then they could be included if (a) administrative costs to the physician of doing so are not excessive and (b) it is anticipated that putting the doctor at risk for the cost of those services will lead him or her to significantly reduce the extent of ordering.

The discretionary services, which could be provided by the principal physician or could be referred out, are obvious candidates for inclusion. Excluding consultations, but including own-provided services will lead to excessive substitution.³ At least some of the cost of such substitutable referral services should be included in the bundle.

One strategy for doing so which would not require principal physicians to pay consultants would be to include the cost of consultant services, and perhaps diagnostic tests, as complete or partial offsets against the fee that Medicare pays to the principal physician. This avoids the need for the doctor to arrange for fee splitting. Medicare would pay directly for the diagnostic tests and consultations and then charge the principal physician based on the level of expenditures for such services.

The configuration of such a penalty can be calibrated to focus on behavior that is likely to be inappropriate. For instance, charges could be imposed for consultations in excess of some fixed number, and even increased at an increasing rate for larger deviations.

This approach could be used both with services provided by others and those provided by the principal physician. Suppose that Medicare typically pays \$40 for a follow-up visit, and suppose that on an average a procedure is accompanied by two follow-up visits. One could reduce the payment for the procedure by say, \$20, for each follow-up visit in excess of two, or else simply pay only \$20 for visits in excess of two. The former strategy would make more sense for services which are sometimes performed by the principal

³ How much substitution would depend on a comparison of the marginal profit for own-provided services under the old fee-for-service system with the marginal cost of those services, which is the lost profit under a system which bundles all own-provided services.

physician and sometimes by others. Or Medicare might pay less than full price for such services when combined with a primary procedure, while paying full price for the same test or consultation visit if given in isolation.

Does such a system of penalty charges have as much theoretical appeal as capitation payments? At one level, the answer is "no". Since the number of Medicare beneficiaries in the next time period can be predicted with a sizeable degree of accuracy, it would be easier to predict total expenditures -- which will be fixed whether the incentives work or not. The major argument for such a bundling arrangement is that, compared to capitation, it requires less institutional restructuring. It might also be subject to less danger of preferred risk selection, though this is far from certain.

It should be noted here that, while one intent of bundling is to decrease the incentive to induce demand for complementary services, bundling could potentially decrease physician willingness to provide beneficial services. Presently, surgeons are often criticized for not providing good pre-and post-operative care, as are internists for not counseling patients or performing thorough histories. At least part of the their reluctance to provide these services is due to relatively low payments for them. Bundling could lead to an even greater reluctance to provide services that are complements to the principal service, and lead to adverse effects on quality. While this is a potential hazard of any volume control, the incentive to simply not provide a (complementary) service, rather than to provide it more judiciously, is somewhat stronger with this control.

5.7.3. Summary

Bundled payments do involve less control over total spending than does full capitation. Not only that, precisely because some services are left out of the bundle, there will be an incentive to substitute those services, when possible, for ones included in the bundle.

These theoretical defects might be offset by greater feasibility in terms of institutional structure. If arranging that structure is left to physicians, this may not turn out to be true. There are some physicians who would have an easy time of it, such as those already practicing in multi-specialty group practices. But most physicians are unlikely to be eager to arrange to share fees with suppliers of other inputs to the bundled treatment.

One possibility is for "increased bundling" to be left as an alternative that a doctor could elect. In return for a payment less than current average Medicare payments, a doctor willing to serve as principal physician could receive a bundled payment and would be able to gain by choosing a more economical mix. To guard against preferred risk selection, it would probably be necessary to require that this option be selected on a participating basis, rather than patient-by-patient.

Appendix: Theoretical Effects of
Bundling Under Behavioral Models

Appendix: Theoretical Effects of Bundling Under Behavioral Models

Bundling Under Profit Maximization

Under fee-for-service a profit maximizing physician maximizes his revenue per encounter. He selects a set of services $S_1 \dots S_n$ that maximizes revenue and minimizes variable costs. The model assumes short term demand is infinite. At the same time the profit maximizing physician would want to ensure that over the long term he has a sufficient number of patient encounters from which to generate the revenue.

Bundling implies that a new unit of payment is defined. The unit of payment may be defined in various ways but would necessarily include a major service. The payment for the bundle will include the major services and some additional services. The price for the bundle will be a fixed price.

We will assume initially that:

All services in the bundle are provided by the physician and that the price paid for the bundle equals total revenue at the fee-for-service rate. Thus the bundle is revenue neutral.

Once the bundle has been defined and its price fixed, the profit maximizing physician will have an opportunity to increase profits by eliminating some of the services previously included in the bundle. In effect he or she will maximize his profits by minimizing the services provided in the bundle. The propensity to minimize provision of services will have a lower bound determined by patient satisfaction, which determines long term demand. Under bundling long term demand is of special importance especially if bundles are defined by encounters. Thus the number of encounters will determine revenue.

The implications of bundling for profit maximizations are:

1. Expenditures will remain fixed.
2. Volume of services will decline.
3. Quality of care may actually improve if under fee-for-services overutilization resulted in unnecessary services which had harmful consequences. In order to bill for the bundle, the physician will still have to perform the major service but some complementary and potentially harmful services will not be provided. It is also possible that substitute services, which are less profitable but are at least as effective as the services provided under FFS, will be provided. It is, however, possible that services with positive contribution to quality will not be provided when needed. Such a phenomena will reduce quality. However, long term considerations will limit the reduction in quality.

4. We assume now that some of the services included in the bundle are referred and the physician is to split his fee with other providers.

A profit maximizing physician will attempt to reduce the number of referrals and pocket the referral fee, thereby reducing his costs. As argued in section 5.7.2. (above), if referrals are considered a necessary part of the treatment they should not be included in the bundle and they should be billed separately. Referrals, used under fee-for-service, which are judged unnecessary should be included in the bundle.

5. The bundle includes services which could be performed by the physician or referred.

The analysis here, in terms of the physician incentive does not differ from case 2. The problem however is to ensure that "necessary referable" services are actually provided. To ensure provision, these services should not be part of the bundle. Separate billing should be allowed. If the physician is allowed to bill for these services, he or she would provide the same amount as under fee-for-service. If the physician is not allowed to bill for this service he or she will be indifferent and refer the same number as he or she performed under fee-for-service.

The policy implication of this analysis is that bundling should be implemented in conjunction with a monitoring system which ensures that services paid for are actually provided and that quality of care is not deteriorating because of changes in practice patterns.

Bundling Under Income Targeting

In this model further demand inducement is possible. Therefore we have two conditions to investigate: comprehensive bundling and partial bundling. Partial bundling refers to the situation where some major services have been bundled but a majority of services have not. This is the more likely scenario for Part B. Comprehensive bundling implies that all/or a majority of services have been included in bundles.

Partial Bundling

1. All services in the bundle are provided by the physician.

If the bundle defined is revenue neutral, then the income targeting physician will not change his behavior. The bundle reflects his professional judgment and satisfies his net income requirements.

However, bundling implies standardization. It must therefore be the case that for some physicians the bundle is not net income neutral. If the bundle does not include services a physician currently provides, he needs to make a choice between his professional subjective valuation and the possible loss of

revenue if he persist in providing the service (assuming no balance billing). If balance billing is assumed, he will impose out-of-pocket expenses on his patients, which will have negative implications on the long term demand for his services. The physician will most likely decide to provide the bundle for which he is paid. Whether the physician continues to provide services as he did before bundling or reduces the services provided, bundling will result in a loss of net income.

The physician will attempt to induce demand, in order to obtain his target income. In a world of partial bundling the induced demand will be directed towards services not bundled. If those services produce lower net income their volume will increase by more than the reduction in volume of the services which produced higher net income. Expenditure may therefore increase, depending on the reimbursement rates of the services not included in the bundle and the services omitted from the bundle.

Consequently depending on the way bundles are defined we may observe reduction in the volume of some services excluded from the bundle. However, induced demand will result in an increase in the volume of other services. The net impact on expenditure would also depend on the physician's willingness to induce demand at levels higher than before bundling.

Under conditions described here quality of care may decline. Two factors may contribute to a decline in quality. First, if items contributing to quality are omitted from the bundle because of the underpricing, services not included in the bundle should be provided. Second, to the extent that induced demand occurs quality may decline.

Note that the analysis regarding the variation of services around a bundle will apply to a strategy in which the price for the bundle is lower than the revenue under fee-for-service. In this case physicians will need to supplement their income. The impact on total expenditure, volume and quality would be the same but at a larger magnitude.

The Bundle Includes Some Referred Services

In this model, as long as the bundle is revenue neutral and allowance is made for the necessary referrals, physicians are given an opportunity to practice in accordance with their subjective professional judgement and practice will not change, since income requirements are satisfied. However, if large variation exists across the bundle definition it might be better to exclude the referred services from the bundle and bill them separately. If the referred services are high cost items then one may require, in addition to separate billing, a second opinion or any other pre-utilization approval. Such a measure would avoid creation of additional induced demand and provide the physician with an opportunity to provide care in a way consistent with his professional subjective judgement.

The Bundle Includes Services Which Can be Performed by Physicians or Referred

If the bundle is net income neutral, no change will be observed. The referred services should be included in the bundle. Physicians will choose to provide them, if they did so under fee-for-service. If they used to refer the service they could still do so. If bundles are underpriced, referrals will decline. Referred services may still be provided by the physician, but the volume of some other services will decline.

Comprehensive Bundling

If a comprehensive bundling system is developed, and physicians lost income how are they going to induce demand? If the bundles are defined by encounters they will have to create more encounters with their pool of patients. Alternatively, they may give more prominence to secondary diagnosis and symptoms and modify the case mix to create new bundles. Policies to prevent such adjustments need to be developed and implemented as bundling is introduced.

6.0. SURVEYS OF MEDICARE AND PRIVATE SECTOR CARRIERS

6.1. Methods

The purpose of this section is to describe the range of approaches currently used by all Medicare Part B carriers and a sample of private health insurance carriers to control volume and intensity. Two surveys were conducted to assess these approaches. First, in consultation with the Bureau of Program Operations (BPO) of HCFA, the University of Minnesota designed a survey of Medicare Part B carriers (a copy of the survey is attached as an appendix to this report). An instruction page explained that the purpose of the survey was to analyze the effectiveness of methods currently used to ensure that Medicare payments are made only for medically necessary physicians' services. A cover letter from the Director of the Office of Program Operations Procedures (OPOP) urged carriers to cooperate with the survey.

The second survey, of private insurance carriers, was designed by the Health Insurance Association of America (HIAA), with the advice of the University of Minnesota. Questions related to the methods used by private carriers to control the volume and intensity of physician services. Each carrier was asked about its commercial and group health insurance business in three areas: commercial insurance products, PPOs, and HMOs. Copies of the survey instruments are appended to this report.

The survey of Medicare carriers was mailed on September 12, 1988, to 46 carriers. It was followed, about 2 weeks later, by a telephone call from the University of Minnesota (carriers had been notified of this call by the instruction sheet that accompanied the mail questionnaire). The purpose of this call was to review any problem areas and to write down the answers to the survey questions. Carriers were then instructed to return the survey by mail. This method -- utilizing both mail surveys and telephone follow-up calls -- has been found to be an accurate method for collecting complex information such as that requested by the carrier survey. By early November, 44 surveys had been returned for a response rate of 96 percent.

The HIAA survey was conducted in June and July, 1988, by trained telephone interviewers from HIAA's Washington office. One hundred twenty companies were selected from a sampling universe of 132 companies. A response rate of 100 percent was achieved to this survey.

Both surveys contained "closed" and "open" ended questions regarding carriers' programs to control volume and intensity. Closed-ended questions present the respondent with a limited number of categories from which to select an answer; alternatively, they may ask for a numerical or qualitative rating of the effectiveness of a volume/intensity control. Open-ended questions offer the respondent more latitude in his/her answer (e.g., what is your overall feeling toward the effectiveness of a control?). This type of question is often useful for understanding the context in which controls are used; however, it may be difficult to categorize and compare responses to open-ended questions.

6.2. Results

6.2.1. Medicare Carriers

We will begin by presenting the results for medical review (MR) activities of Medicare carriers. Medical review screens are manual or automated edits designed to suspend processing of services meeting specified selection criteria. Services are then evaluated to determine whether they were medically necessary. In general, medical review screens may be of two types: prepayment and postpayment. These screens are distinguished by whether selection and review occurs before or after the claim has been paid.

Until recently there was only one mandated prepayment screen: claims from physicians making multiple visits to nursing home patients were automatically reviewed. However, 13 national screens, all dealing with volume and intensity of physicians' services, are now mandated by HCFA. Claims falling above certain parameters, e.g., more than 31 hospital visits per three months, are reviewed for medical necessity before they are paid. The carrier (typically utilizing a nurse, physician, or clerk) reviews the claims against criteria of medical necessity. Claims deemed unnecessary may be denied or reduced. Carriers may set tighter parameters for screening than those mandated by HCFA and there are numerous local screens in addition to the 13 nationally mandated screens.

Table 1 shows how many carriers (out of 44) implemented the screen before the mandated date, and how many use a screening parameter tighter than the mandated parameter:

Table 1

Information on Nationally Mandated Prepayment Screens

Screen	Mandated Date (mo/year)	# of Carriers With Prior Date	Mandated Parameter	# of Carriers With Tighter Parameter
Routine foot care	10/84	6	1 treatment per 60 days	3
Mycotic nails	10/84	11	1 treatment per 60 days	2
Nursing home visits	10/84	16	1 visit per month	1
New patient office visits	10/84	7	1 comprehensive physical exam per carrier history period	3
Holter and real-time monitoring	10/84	6	1 instance per 6 months	0
Chiropractic	1/86	32	12 spinal manipulations per year	8
Concurrent care	1/86	34	1 doctor of same specialty billing for in-hospital services on same day	3
Hospital visits	1/86	29	31 times in 3 months	5
Comprehensive office visits	1/86	28	1 per 6 months	2
SNF visits	1/86	33	2 subsequent care visits in 1st week, 1 visit per week thereafter	6
Injections	1/86	30	24 per year	4
Urological supplies	1/86	28	2 catheters per month	1
Replacement of post-cataract external prosthetic contact lens	1/86	27*	1 per eye per year	2

*2 missing values

This table shows that, although many carriers implemented the screens before they were mandated, few of them use tighter screening parameters. In fact, we had planned to tabulate the values of all tighter parameters but gave this up because there were so few cases. One explanation for why carriers do not exceed the mandated parameters is carrier ignorance. We discovered that most of the screens implemented prior to the mandated date used tighter parameters. However, when the screen was mandated, carriers tended to make their systems conform with the federal mandate. Thus, many of them relaxed their screening parameters. They apparently thought that HCFA would not allow tighter parameters.

Eight carriers use tighter screening parameters for chiropractic services. On close examination, however, most of these carriers prohibit more than 12 spinal manipulations per year or 1 per month. Their extra screen will catch suppliers who perform 6 spinal manipulations within 3 months, for example. Two carriers have implemented very strict screens: one reviews all chiropractic claims for medical necessity, and one requires x-rays to substantiate the claims.

Carriers were asked to describe any problems they encountered in implementing mandated prepayment screens. Responses to this open-ended question indicated a general increase in complaints about government interference in medical practice, but no significant problems. Several carriers offered constructive suggestions about how to implement a new screen. They suggested that it should be publicized in advance (through newsletters, etc.). One noted that the implementation date is effective immediately and gives little time to instruct the providers on the new screens. Once the provider is notified and edits/audits are in place, no major problems are encountered. Interestingly, most of the complaints about screens seem to come from podiatrists and chiropractors. One carrier commented, "The chiropractors are very organized and they continually ask for further reviews."

Carriers were asked to name the skill level of the person who normally performs medical review for mandated screens. In 27 cases this was a claims examiner with limited medical training. Nurses were mentioned 12 times, and physicians were mentioned 5 times. Many carriers also mentioned a second skill level, although they were not required to so. The most frequently mentioned second-level claims reviewer was a nurse.

Carriers were given the opportunity to "sound off" by naming up to three mandated screens that are not cost-effective, in their opinion. Only 6 carriers failed to name at least one ineffective screen. Table 2 shows the number of times each mandated screen was cited as ineffective:

Table 2

Carriers' Opinions About Ineffective Mandated Screens

Screen	1st Mention	2nd Mention	3rd Mention	Total
Routine foot care	11	4	1	16
Mycotic nails	0	3	2	5
Nursing home visits	3	0	0	3
New patient office visits	0	1	0	1
Holter and real-time monitoring	0	1	3	4
Chiropractic	0	1	3	4
Concurrent care	0	0	1	1
Hospital visits	6	1	1	8
Comprehensive office visits	1	2	0	3
SNF visits	2	1	2	5
Injections	12	4	3	19
Urological supplies	1	7	2	10
Replacement of post-cataract external prosthetic contact lens	2	7	3	12
None mentioned	6	12	23	41

Several interesting features are shown by Table 2. First, carriers consistently rate certain screens as ineffective. Routine foot care and injections top this list, followed by urological supplies and post-cataract contact lenses. Carriers had previously stated that the foot care screen was difficult to implement. However, only 4 carriers view chiropractic prepayment screening as ineffective, although this was the most likely screen to have implementation problems. Finally, almost all carriers rate the following screens as cost-effective: nursing home visits, new patient office visits, concurrent care, and comprehensive office visits.

We previously noted that carriers use numerous local screens in addition to the 13 mandated screens. We asked each carrier how many such screens it used. The median response was 21, and one carrier had 121 local screens. Apparently, carriers are quite active in this area. Carriers were asked if they would recommend up to 3 of these screens to include in the list of nationally mandated screens. Although the responses to this question were extremely varied, several regular patterns appear. First among these is the carriers' opinion that consultations are being abused. This was mentioned about 20 times by the carriers. This was followed closely by hospital visits. Carriers were concerned both about the number of visits per hospitalization and the number per time period. Third, carriers recommended screening office visits more carefully than is currently mandated. Finally, hospital emergency room visits were cited as an area to review.

Interesting, abuses of particular services or procedures were not frequently mentioned by the carriers. Several exceptions to this rule were: multi-channel lab studies, vascular studies and venipuncture, critical care, seat lifts, and ambulance services.

Postpayment medical review (MR) is a process by which practice patterns of physicians/suppliers are compared to statistical norms on them and their specialty groups. The general purpose of this process is to identify suppliers whose practice patterns depart from recognized norms, and to correct overutilization of services by recovery of overpayments, postpayment monitoring and education.

Knowledge of service area data is one of the most important requirements in operating a postpayment screen. Carriers, therefore, are directed to "use acceptable statistical techniques in your postpayment utilization safeguard system to array data by specialty group." Once the data have been analyzed, carriers are directed to "use the appropriate statistical tool which will best identify the physician/supplier who needs further investigation by the postpayment MR staff." (Citations are from the Medicare Carriers Manual, Part 3-Claims Process, June, 1988, p. 7-141.)

Carriers are required to make postpayment pattern-of-practice comparisons for the following service categories: office visits, home visits, hospital visits, SNF (skilled nursing facility) visits, injections, EKGs, surgery, office lab services, office diagnostic x-ray, and physical therapy. There are two peer practice pattern comparison ratios to calculate for each service: Ratio I is the number of services provided by a physician/supplier per 100 beneficiaries; Ratio II is the number of services per beneficiary who actually

received services in that category. After calculating these ratios, carriers are instructed to determine the point at which the ratio exceeds the norm. They may use percentiles, index values, medians, modes, etc.

The Carrier Survey requested each Medicare Part B carrier to furnish the critical value of Ratios I and II that would cause a supplier to be selected for postpayment review. However, the carriers responded that "critical values" could not be determined by simple formulas. Instead, more-complex methods were used. These typically involved selecting a point on the distribution of ratios. In 14 cases the point selected was 2 standard deviations above the mean. Three carriers used a standard deviation other than 2.0, and two used 2.0 standard deviations plus the top 150 physicians.

Three carriers automatically review the top physicians within each category, whether or not these physicians have unusual practice patterns. The chosen cutoffs vary widely: 20, 200, and 300 physicians are selected for review. Four carriers used percentile cutoffs, reviewing the top 25, 20, 5, and 3 percent of all physicians, respectively. Seven carriers used a specified percentage above the mean to select physicians for review. Both of these methods are similar to using a standard deviation-based cutoff, although this would vary widely (for example, reviewing the top 5 percent of all physicians is similar to using a 1.65 standard deviation cutoff, but reviewing the top 25 percent is similar to using a much lower cutoff -- about 2/3 of a standard deviation above the mean). Six carriers had no specific standards for postpayment review. Two did not answer this question; one gave an answer that could not be classified; and only two used a specified number for each ratio.

Consequently, it appears that carriers use widely different postpayment standards. Although these differences may reflect optimal adaptations to local environments, it is also possible that some of the standards lack a carefully justified basis. This is especially likely for those that are based on reviewing a predetermined number of physicians. Unless this system is based on an implicit recognition that these physicians have unusual practice patterns, it is difficult to argue that a simple numerical formula, e.g., "review the top 20 physicians," is a well-designed postpayment MR strategy.

As in the case of prepayment review, each carrier was asked to name the skill level of the person who normally performs postpayment medical review. Nurses were mentioned 24 times, followed by claims examiner with limited medical training, mentioned 12 times. These responses clearly indicate that postpayment MR is a more highly skilled function than prepayment review. Fourteen carriers mentioned a second person who performs postpayment MR; this was usually a physician or nurse.

As we have discussed in Section 5.6, collapsed coding may be a desirable method for controlling volume and expenditures, although our theoretical discussion also indicates that the effect of coding changes may be ambiguous. In practice, how many Medicare carriers have implemented coding system changes, and what is the opinion of those who have implemented them regarding their effectiveness? Each carrier was asked if it has changed the coding of physician services to prevent physicians from "upcoding" procedures (e.g., has

the carrier reduced the number of office visit codes it recognizes for payment purposes?).

To our surprise, carriers placed a different interpretation on this seemingly straightforward question. Many responded that they had implemented coding system changes, but what they described was really a downcoding screen. For example, if a physician submitted two claims for initial office visits from the same patient, the second claim was automatically downcoded to a follow-up visit. When the interviewer attempted to clarify this question, carriers responded that they were required, by law, to recognize HCPCS codes on the claims. Their interpretation of this requirement was that HCPCS codes must also be recognized for billing. It did not appear to them that two codes could be combined or collapsed for billing, e.g., that initial and follow-up office visits could be paid at a single blended rate.

Only one carrier appeared to have a coding system program that went beyond the downcoding screen. This carrier's unique approach is described verbatim:

"Although codes were not changed to prevent upcoding, many of the allowances are the same. This occurred when we converted to HCPCS with pricing being equal for many levels contained in CPT-4. Therefore, upcoding is somewhat controlled by the pricing and payments."

Unfortunately, this carrier has not formally evaluated the effectiveness of its coding program. Thus, no evidence is available by which to judge the effect of the coding changes on cost or utilization.

Next, carriers were asked if they extensively "bundle" physician services into broader categories (e.g., global fees for surgery). Most of the carriers have such programs for surgery. In one case, the carrier remarked sarcastically that global fees for surgery have been used in its state "since the beginning of time." Global surgical fees were applied in different ways by the carriers, some of whom appear to be extremely creative in this regard. To quote one carrier:

"... policy calls for bundling of ancillary services as well as pre and post operative care. There are approximately 700 ancillary combinations. One of our most common combinations is physical therapy treatment to one area plus two or more modalities (e.g., whirlpool, paraffin bath)..."

Another carrier achieved global surgical fees by automatically denying pre and post operative visits, diagnostics, and incidental surgical procedures.

Despite their use of global fees for surgery, few carriers have extended the bundling concept to other services. Responses to an open-ended question revealed these exceptions: three carriers had a single fee for automated lab panels; and two carriers applied bundling to all bills for the same month, place of service, type of service, procedure code and provider number (except

for pathology, anesthesiology and psychiatric services that have specific lumping guidelines).

The last question on the Carrier Survey gave each carrier the opportunity to "Describe other techniques for controlling volume and intensity that you have successfully used under Part B of Medicare." Responses to this open-ended question were read and analyzed. The most frequently mentioned activity was physician education. This often took the form of simple newsletters. More comprehensive educational programs involve the carrier's provider relations department in telephone contacts or on-site visits to the doctors. Although the effectiveness of educational programs is difficult to evaluate, several carriers mentioned high rates of physician compliance and one even cited a "savings" of almost \$1 million.

It appears that education is most effective if conducted in a non-adversarial fashion. One carrier described the following approach:

"The P.R. (provider relations) department presents monthly seminars to educate providers/suppliers. Quarterly bulletins are sent. We have the ability to print messages at the bottom of EOMB's. The post payment unit presents documentation seminars at hospitals and for physician groups. We work closely with our Medical Director to keep abreast of the current patterns in medical practice. Physicians/suppliers who present patterns of overutilization can be put on a prepayment review to better control his (sic) volume of work processed and collect data to educate him."

Similar comments could be made about the educational programs used by other carriers. This non-adversarial approach supports our analysis of clinical guidelines and professional education (see Section 5.2.), which asserted that guidelines and education are more likely to be effective when they are clearly defined, data based, professionally derived, and emanate from respected sources.

Another general conclusion to be drawn from these open-ended responses is that carriers know who the "bad doctors" are. Several mentioned that postpayment review has identified individual physicians who present patterns of upcoding and overutilization. Specific screens may be devised for these providers. For example, one carrier automatically denied specified procedures when billed by a designated physician. Another noted that postpayment review has identified several physicians having patterns of upcoding. Repayment screens have been developed for these individuals which reduce the level of payment if documentation for the higher code is not provided. Finally, the carriers claim to have aggressive fraud and abuse programs for physicians who habitually abuse the Medicare program.

6.2.2. Survey of Commercial Health Insurance Carriers

This section of the report highlights findings from the Health Insurance Association of America's (HIAA) Managed Care Survey conducted in the summer of 1988. HIAA represents 194 dues-paying organizations that write 85 percent of

the commercial health insurance sold in the nation. The objective of the survey was to document the mechanisms that commercial insurers use to control the rising cost of health care services. HIAA researchers conducted a similar survey during the summer of 1986, making it possible to document recent changes in cost control efforts during the past two years, as well as taking a snapshot of current practices.

Methods

In the spring of 1988, HIAA researchers sent a letter to the executive responsible for the group, individual, preferred provider organization (PPO) and health maintenance organization (HMO) lines of business. The letter described the survey, listed specific questions, and indicated that a HIAA researcher would be calling shortly.

The sample included 123*of the 194 dues-paying organizations. The sampling frame was stratified according to prior information about the lines of business each insurer had entered. All insurers known to have entered the HMO and PPO markets were surveyed. Survey results can be presented in two formats. One approach is to show what a typical insurer is doing. Here, the unit of observation is the individual insurer -- the number of respondents to the HIAA survey is weighted to equal the number of insurers in the dues-paying universe. The second format is to present developments for the commercial insurance industry as a whole. In this case, each insurer is weighted by its relevant volume of business. Thus, large insurers count considerably more than small insurers in this presentation.

HIAA researchers interviewed a maximum of 4 executives with each insurer (group, individual, PPO and HMO executives). The typical interview lasted 15 to 25 minutes. One hundred percent of the interviews were successfully completed. This may be attributed to the persistence of the interviewers, and to the interest of HIAA's members in the survey. Item response rates exceeded 90 percent for most questions.

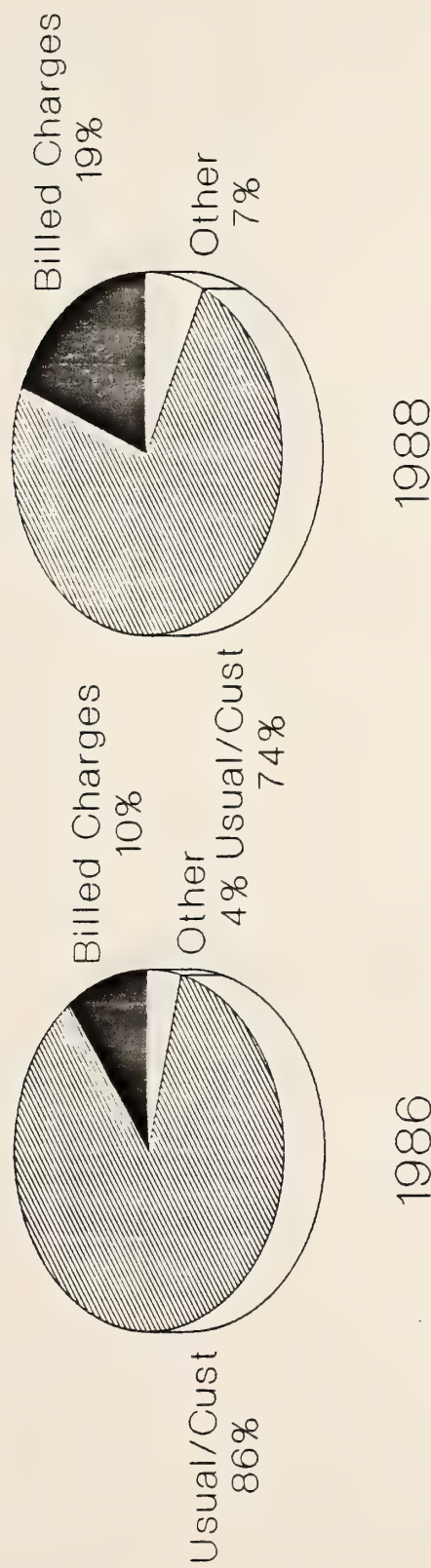
Survey Findings

Conventional Group and Individual Insurance

Group and individual insurers reimburse physicians largely on the basis of usual and customary charges. Figures 1 and 2 (both based on percentage of business) show that UCR covers 74 percent of conventional insurance business and 56 percent of individual business. For group insurers there has been a slight movement toward billed charges during the past two years.

When the company was the unit of observation, HIAA found that UCR was utilized by 111 conventional insurers. The next most popular reimbursement method was billed charges, utilized by 12 companies. Only 6 respondents utilized "controlled" methods (discounted charges, fee schedule, or capitation) and 3 companies utilized "other" methods of paying physicians as their most common reimbursement method.

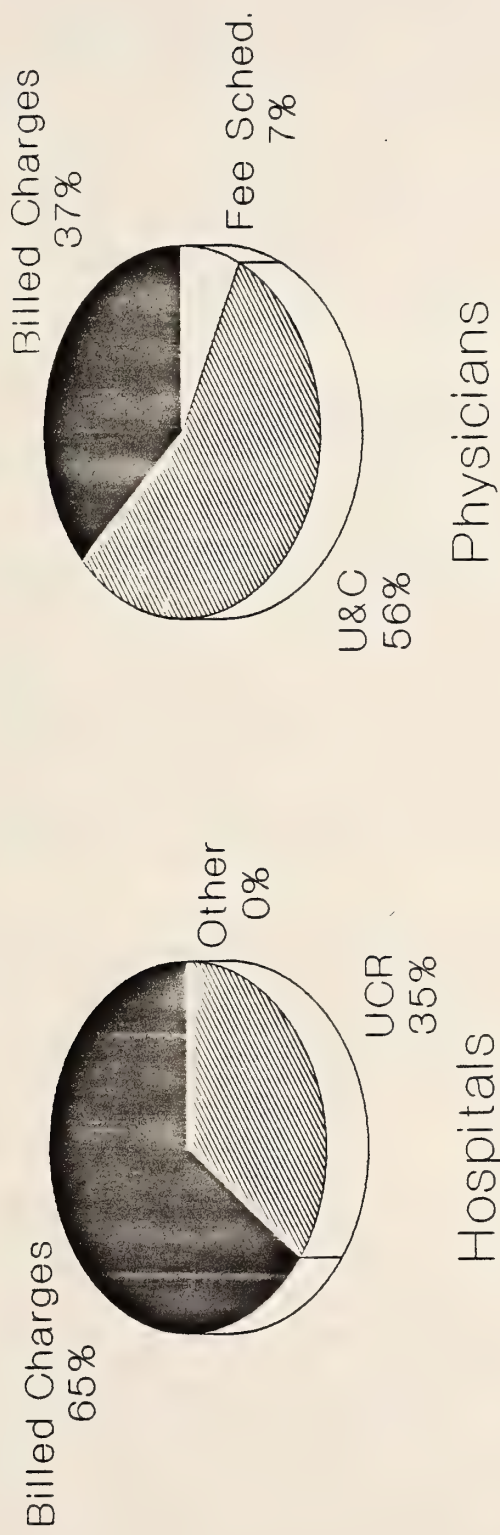
How Did Conventional Insurers Pay Physicians For Their Services In 1986 And 1988?



Sources: HIAA Managed Care Surveys
1986 and 1988

Figure 1

How Do Insurers Pay Hospitals And Doctors Under Individual Policies? (1988)



Source: HIAA Managed Care Survey, 1988

Figure 2

Utilization review (UR) is clearly the most popular control over hospital utilization in conventional health insurance. Most companies offer more than one type of utilization review in their group business. For example, 120 companies provide pre-admission certification, 109 provide concurrent review, and 85 provide retrospective review of inpatient care in their group business.

Utilization review grew at a rapid pace between 1986 and 1988 in group business, more than doubling the percent of business covered (Figure 3). However, UR focuses on keeping patients out of hospitals rather than controlling inappropriate care in ambulatory settings. Only 9 percent of business uses physician profiles. The individual business market is now at the "takeoff" stage with regard to utilization review (Figure 4).

Of the 69 companies that have formally evaluated the effectiveness of pre-admission certification, 56 believe that it reduces total cost of group insurance. Only one company reported an increase in total cost; the others were either undecided or reported that they "don't know". The average decrease in total cost was estimated to be 7 percent by 52 companies. Companies that had not formally evaluated their pre-admission certification were asked about their "overall feeling" toward the effectiveness of pre-admission certification. The most popular response was "somewhat effective," although some companies felt that the program was effective only when it began. Similar opinions were voiced about the effectiveness of concurrent review. Only one company had a negative opinion of its concurrent review program, based on a formal evaluation. Retrospective review, the activity least likely to be evaluated, had a mixed reception from the insurers. Figure 5 summarizes insurers' perceptions regarding the effectiveness of utilization management in controlling claims costs in their group business. Individual insurers are more "bullish" about UR (Figure 6).

Although the companies agreed that utilization review reduced total cost, they gave it poor marks for controlling cost and utilization outside the hospital. Most of the respondents answering this question, believe that pre-admission certification increases cost and utilization outside the hospital. (This question was asked only for insurers that had not formally evaluated their pre-admission review program. Thus, it may not be generalizable to the other insurers.) Concurrent and retrospective review of inpatient care were not seen to have this adverse effect, however.

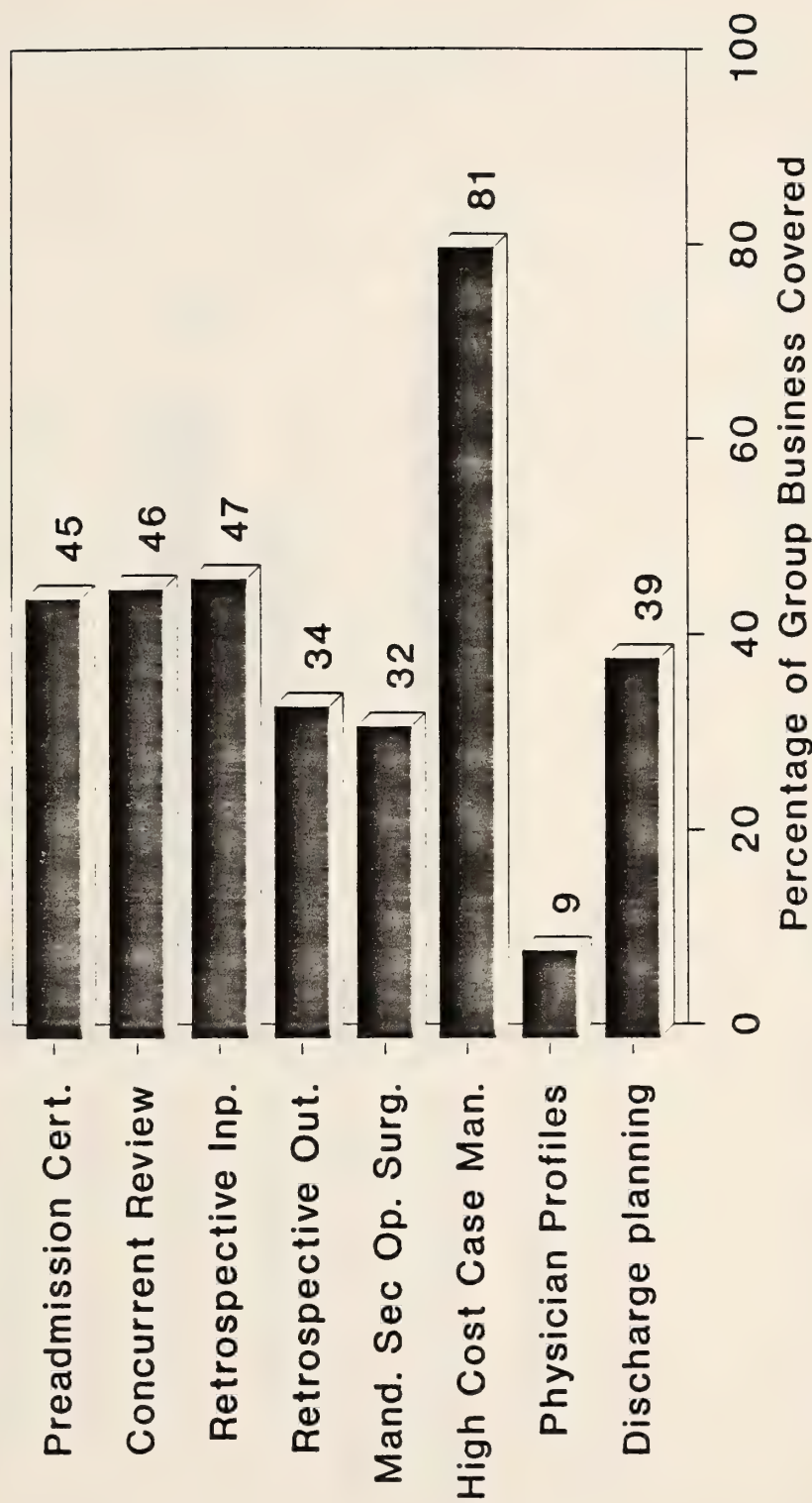
Thirty-six of the insurers practice retrospective review of outpatient care, but only 12 have formally evaluated their program. Seven of these believe the program reduces total cost, versus four undecided insurers and two that report higher total cost. The 13 firms that have not formally evaluated this program gave it more mixed reviews, however. Four of these firms believe that retrospective review results in higher outpatient cost/utilization. This is surprising, since the program is targeted toward outpatient care. It is possible that retrospective review of outpatient care produces some savings, but that these are outweighed by the costs of conducting the program.

Preventing multiple physicians from billing for the same service, such as eliminating payments for assistant surgeon, is the next most popular control,

following utilization review. Thirty-one companies practice some form of this policy. However, only 5 companies have formally evaluated the effectiveness of their programs in this area. When asked for their overall feeling, many respondents felt that preventive multiple billing has reduced cost and utilization outside the hospital.

Percentage Of Group Business Covered By Various Utilization Review Activities (1988)

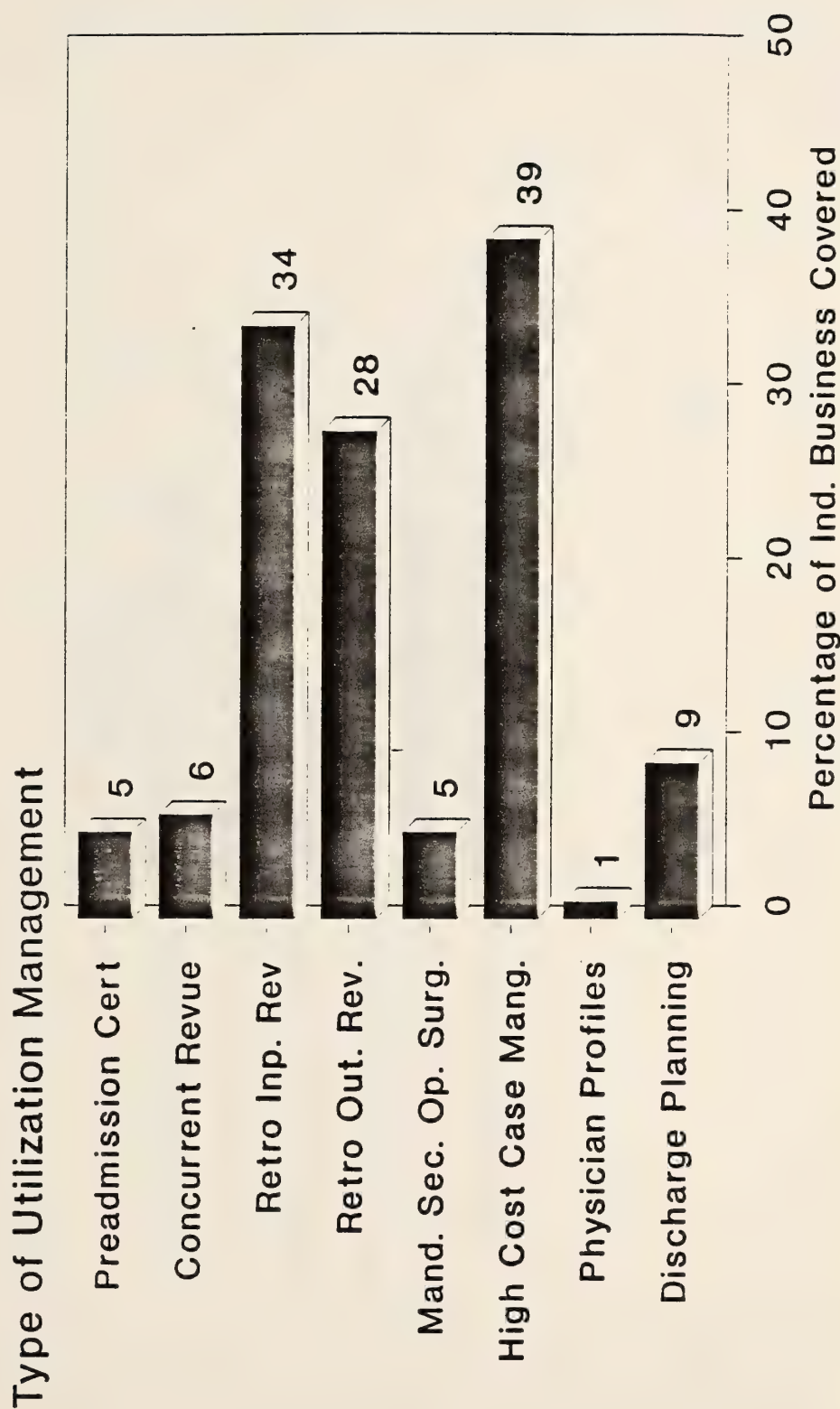
UR Activity



Source: HIAA Managed Care Survey, 1988

Figure 3

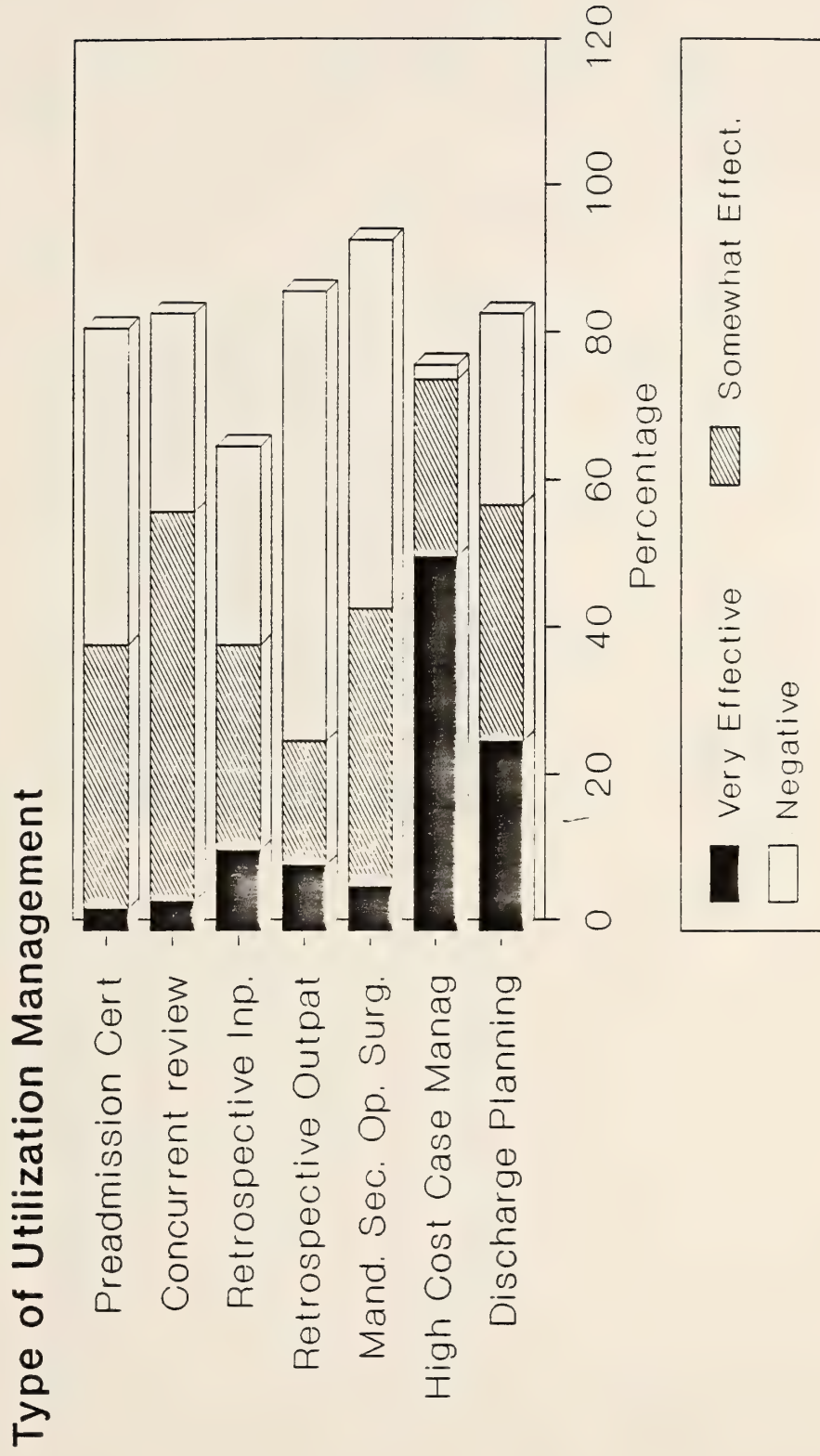
Percentage Of Business Covered By Utilization Management Programs Under Individual Business, 1988



Source: HIAA Managed Care Survey, 1988

Figure 4

Perceived Effectiveness Of Utilization Management In Controlling Claims Costs In Group Business, 1988

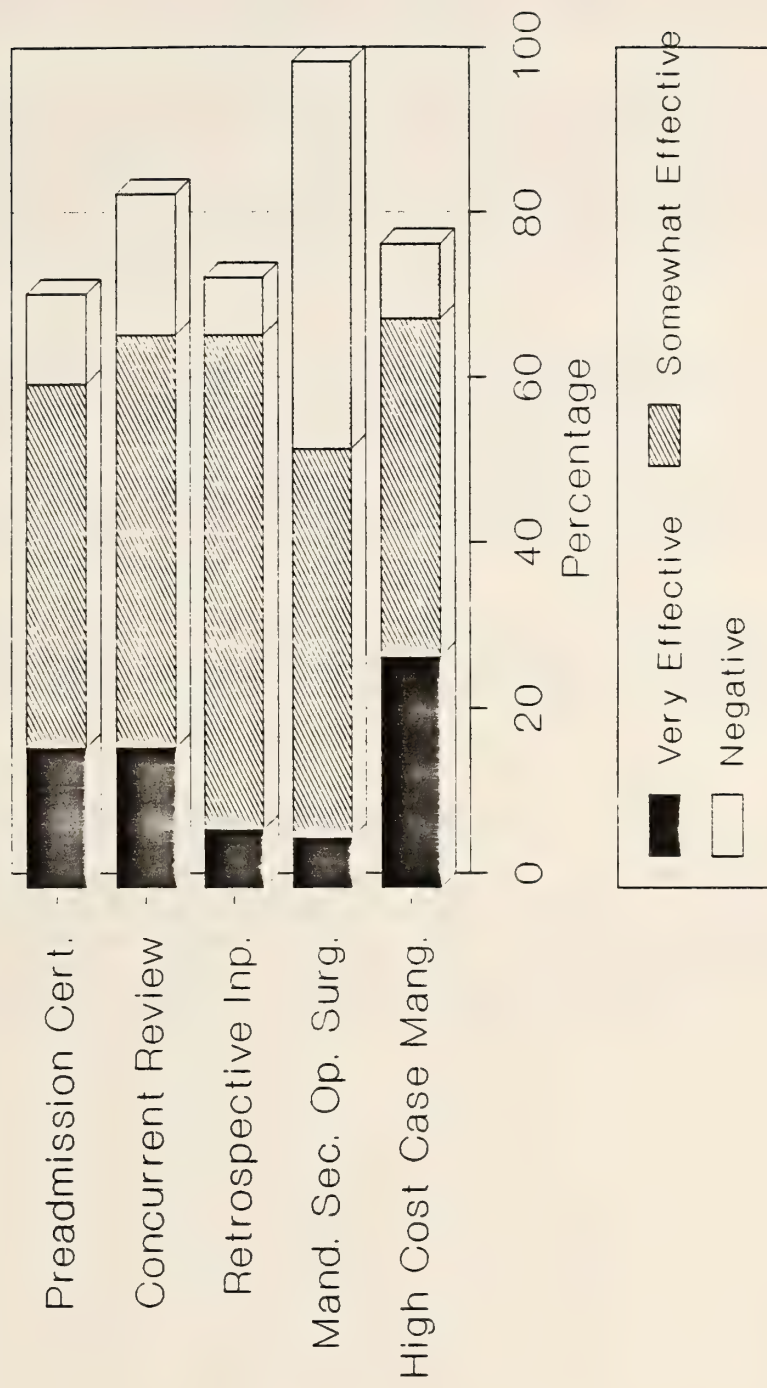


Source: HIAA Managed Care Survey, 1988

Figure 5

Perceived Effectiveness Of Selective Utilization Management In Reducing Claims Costs In Individual Business, '88

Utilization Management



Source: HIAA Managed Care Survey, 1988

Figure 6

Only 13 companies provide physician profiling and feedback in their commercial group health insurance. Companies attempting this form of control believe that it reduces total cost but the sample is too small to place any confidence in the magnitude of estimated cost-savings.

The survey did not ask about the use of clinical guidelines. Although it is tempting to equate physician profiling and feedback with the use of guidelines, it would be incorrect to do so. Profiling can be done without guidelines (the insurance company could simply collect profile data and send it to physicians). Similarly, guidelines do not require the carrier to keep physician profiles. An example is utilization review programs, which can use guidelines without keeping individual physician profiles.

A similar lack of adoption was found regarding coding changes for physician services. Twenty-two companies reported implementing these changes. Many of the respondents were unaware of procedure/coding changes. When this type of volume/intensity control was explained to them, they were generally skeptical that it would control total cost and utilization. Others, when asked why they hadn't opted to change their coding to constrain outlays, responded that "Our company is not big enough," "Just haven't done it," or "We lack the ability to do it correctly."

Finally, 15 companies reported bundling of physician services into broader categories than visits or procedures, such as single payment per episode of illness. The overall feeling toward bundling of services tended to be negative.

Results for Preferred Provider Organizations and HMOs

There were 72 valid cases of firms with PPO arrangements. Unlike conventional insurance, PPOs do not typically use billed or usual charges to reimburse physicians. Forty-seven companies use fee schedules and 42 use discounted charges. When more than one reimbursement method is used, fee schedules are most used often. The following table shows the types of physician reimbursement used by PPOs:

Table 3

Methods Used By PPOs to Reimburse Physicians

Method Used:	# Times Used	# Times Not Used	Don't Know	If More Than One Method is Used Rank According to % of Time			
				1	2	3	4
Billed/Usual Charges	15	56	1	6	4	3	0
Discounted Usual Charges	42	29	1	23	14	3	0
Fee Schedule	47	24	1	38	7	0	0
Capitation	5	66	1	1	1	0	1
Other	3	68	1	0	0	2	0

By volume of business, fee schedules and to a lesser extent discounted usual charges also emerge as the most popular methods used by insurer-sponsored PPOs for paying physicians (Figure 7). There has been a modest movement toward discounted usual charges during the past two years. Insurer-sponsored HMOs use capitation payment as their primary reimbursement method (Figure 8). It should be noted that this is the method for paying the contracted physician group rather than the individual physician.

Fifty-four of the companies responded to this question: "To the best of your knowledge, what percent discount does the PPO receive from the preferred physicians?" The most common answer, reported by 13 firms, was 15 percent. The mean answer was 13.2 percent.

Firms were asked to rank the importance of several methods used by the PPO to achieve its savings. Utilization review, by far received the highest ranking. It was followed by discounts, channeling patients to cost-effective providers, and provider reimbursement methods:

What Were The Primary Methods That Insurer-Sponsored PPOs Used To Reimburse Physicians, 1986 And 1988

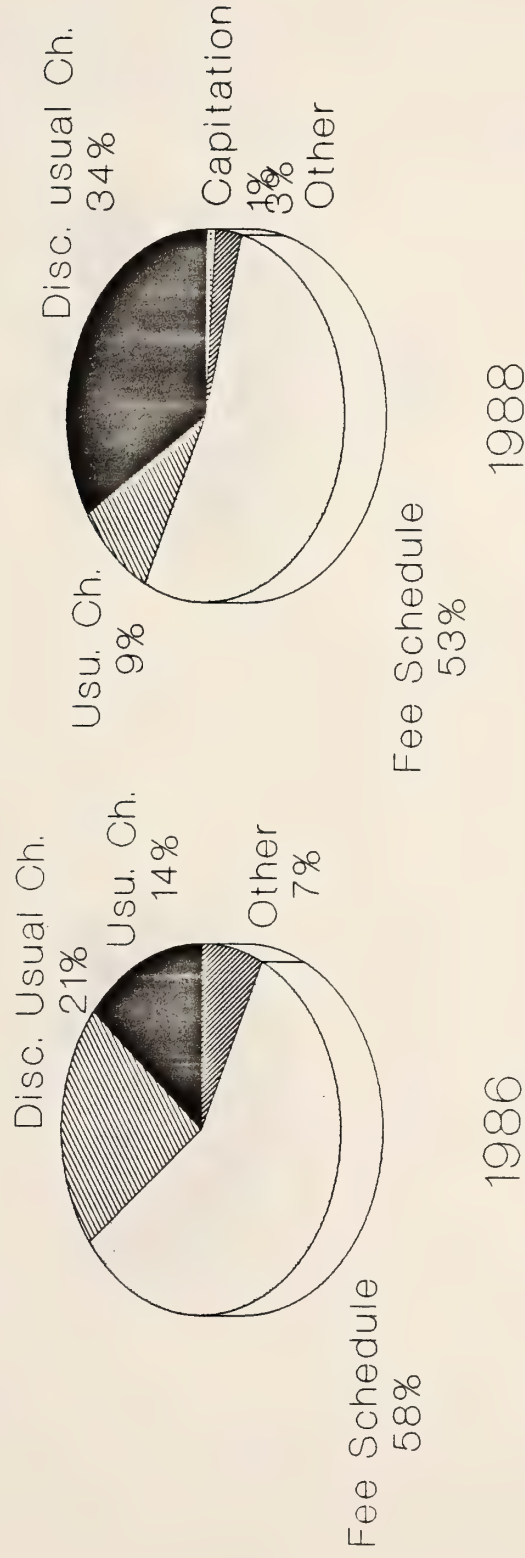
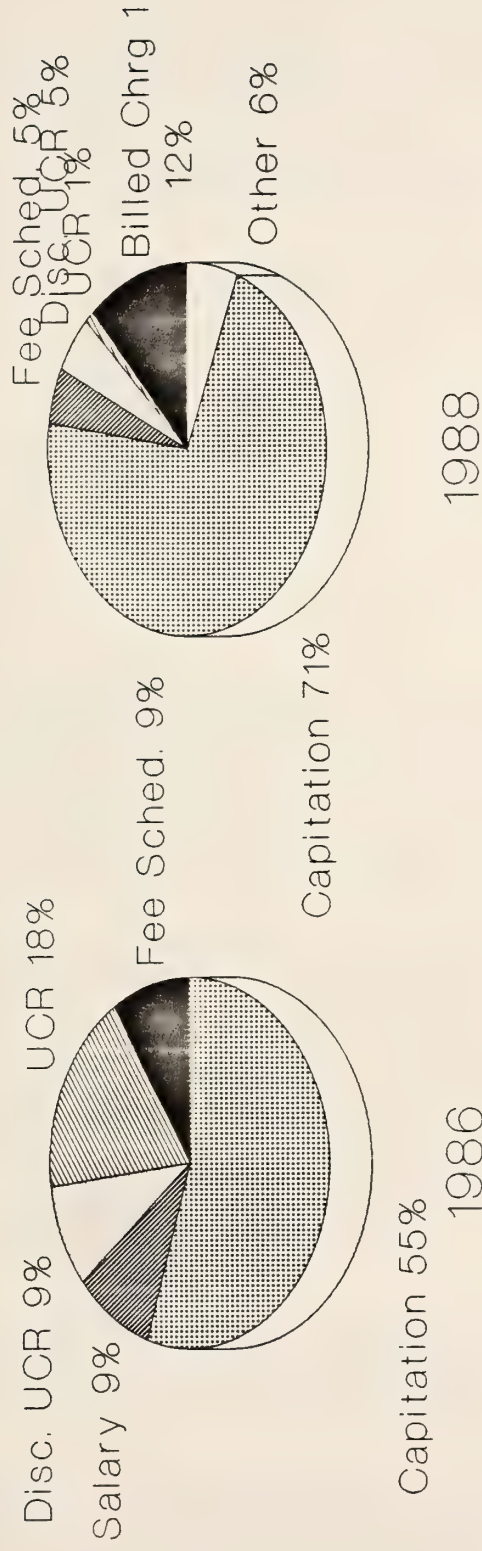


Figure 7

Sources: HIAA Managed Care Surveys, 1988

HMOs Use Capitation Payment For Paying Physicians, 1986 and 1988



Sources: HIAA Managed Care Survey
1986 and 1988

Figure 8

Table 4

Rank of Methods Used By PPOs to Achieve Their Savings

Method	Rank (1=most important, 4=least important)				
	1	2	3	4	Not Asked
Utilization Review	39	10	9	3	11
Discount	14	20	12	17	9
Channeling Patients to Cost-effective Providers	9	19	15	20	9
Provider Reimbursement Methods	5	11	24	21	11

The importance given to utilization review is impressive, especially since PPOs were receiving an average discount of 13.2 percent. Evidently, they believe that utilization review reduces costs by more than 13.2 percent. To an extent, the importance given to "discounts" and "provider reimbursement methods" may measure similar beliefs. Nevertheless, utilization review is still the most important cost-saving method, even if one adds the importance ranks given to discounts and other provider reimbursement methods.

Next, insurers were asked about the utilization review programs conducted by their PPOs. All but two of the PPOs had a pre-admission utilization review program. In 52 cases, utilization review was conducted by an automated system. Somewhat surprisingly, 30 of the PPOs don't levy penalties on physicians or hospitals that refuse to comply with utilization review. Of the 33 companies that mentioned some type of penalty, 13 used suspension or termination of the noncompliant provider, and 14 used a set dollar per episode. Forty-four PPOs penalized patients for failure to comply with the utilization review process. The most common penalties were increased deductibles (mentioned 16 times) and increased coinsurance (mentioned 12 times).

Fifty-one (71 percent) of the PPOs compiled/evaluated physician profiles, up from 40 percent in 1986. Profiles typically are compiled/evaluated by an automated procedure. Twenty-nine PPOs had a distinct psychiatric utilization review program.

PPOs continue to devote less effort in selecting preferred physicians than hospitals. Hospital privileges are the primary criterion (along with a catch-all "other" category) in selecting preferred physicians. About 60 percent of the PPOs identified a data source for selecting preferred physicians. The leading data source was insurer claims.

6.3. Discussion

All Medicare and most private sector health insurance carriers have utilization review programs, although Medicare carriers appear to focus more on reviewing the appropriateness of physician services than do private carriers. This is partly due to the presence of mandated prepayment screens for Medicare. However, many Medicare carriers had these screens before they were mandated and, in some cases, their pre-mandate screening parameters were tighter than those currently used.

Several mandated prepayment screens were cited frequently as ineffective, whereas other optional screens were suggested for inclusion in the list of mandated screens. A systematic study of the effectiveness of prepayment screening is warranted, and such a study is currently underway at the HCFA Research Center.

Medicare carriers suggested that new screens be introduced along with programs to educate physicians about these screens. In general, implementation issues are important and should not be ignored. On the other hand, provider resistance to prepayment screens is not necessarily a sign that they are ineffective.

Medicare carriers operate vigorous postpayment review programs. In many cases, physicians are reviewed if their practice patterns are more than 2 standard deviations from the norm. However, the criteria for postpayment review are not uniform. Standards based on reviewing a prespecified number of physicians may be difficult to justify in terms of detecting unusual practice patterns.

Private insurance carriers were found to operate inpatient utilization review programs that combine pre-admission certification, concurrent review, and retrospective review of hospital inpatient care. The respondents estimate that inpatient utilization review reduces total cost by about 7 percent. Many believe that pre-admission certification increases cost and utilization outside the hospital, however.

For managed fee-for-service programs, utilization review has been largely directed at reducing the use of inappropriate inpatient services. PPOs have made far less effort in selecting preferred physicians than hospitals, often using hospital staff privileges as the major screening criterion. HMOs have historically realized their savings by reducing the number of hospital days.

Private carriers have been slow to innovate in the area of physician payment. All but 8 of the firms surveyed by HIAA still pay usual and customary charges. Only 6 respondents utilized "controlled" methods of paying physicians. Commercial insurers appear to place more emphasis on physician payment reform in their PPO arrangements, which frequently use discounted charges or fee schedules to pay physicians. The average discount was estimated to be about 14 percent. This was exceeded in importance by the cost-saving effect of utilization review.

Perhaps the most significant result from our surveys is that neither public nor private carriers have taken an innovative approach toward bundling physician services and collapsing the codes used for paying physicians. The private respondents, by and large, were ignorant of this concept; Medicare carriers generally thought that they had to recognize HCPCS codes for billing. Medicare carriers appear to be ahead of the private sector in using global fees for surgery, but otherwise they have not attempted to bundle physician services into broader reimbursement packages.

Physician education and feedback is utilized more extensively by the Medicare carriers than by private insurers. This may be due to the fact that they lead the private sector in terms of postpayment review, and this is the area where physician education programs are most likely to occur. Medicare carriers, in general, are sensitive to the need for physician education and information programs.

7.0. DISCUSSION, SUMMARY, AND RECOMMENDATIONS

The continued rise in expenditures by Medicare Part B for physicians' services is due, in large measure, to the increasing volume of services rendered. While the Prospective Payment System seems to have tempered the increase in Part A expenditures, Part B expenditures continue to increase at rates greater than the Consumer Price Index and Part A expenditures. This report is an analysis by the University of Pennsylvania and University of Minnesota evaluating the underlying principles that determine physicians' utilization of medical services and the options available to Medicare to control the volume of Part B services and their attendant costs.

Three models of physician behavior are explored. No one model would be expected to explain fully the behavior of all physicians -- or even the behavior of a single physician -- and the empirical evidence delineating which of the three models would best explain Part B expenditures is lacking. Therefore, while these models may help to provide useful paradigms for predicting the way in which physicians might respond to interventions intended to control volume and expenditures, precise predictions are not possible. It is likely that each of the three models will play some role, but will at best delimit the range of possible responses to these volume control measures.

7.1. Three Models of Physician Behavior

The first model of physician behavior is that of the Clinician as Patient's Agent. In the patient agency model, one assumes that the physician acts as the patient's advocate, and that this is the physician's primary or even only motivation. In this scenario, decisions are made by the physician in the patient's best interest; that is, the physician makes each decision based on what he or she believes the patient would want if the patient had the same information available to the physician. Our formulation of this model includes several components. First, the physician serves as the patient's perfect agent with regard to clinical outcome, thus attempting to optimize the patient's health. The second component of the Patient Agency Model is the physician serving as the patient's economic agent. Physicians become their patients' advocates with regard to the cost of medical care, attempting to avoid undue or avoidable out-of-pocket expenditures for the patient. Each of these components of the patient agency model requires that the physician understand the patient's utilities or values for various clinical outcomes and patients' disutilities for side effects, complications, and inconvenience or discomfort in undergoing a medical service. In addition to being influenced by the patient, the physician as the patient's agent will also be influenced by professional standards and preferred styles of practice. In the best case, these professional standards will hold each physician to a measure of quality expected by the profession, and to standards of professionalism and maintenance of expertise that society would expect of a self-policing and self-controlling profession. Other professional influences may be less idealized and more a function of the sociologic temper of the profession in a given area or specialty.

The role of the physician as agent of the patient may conflict with the second model of physician decision making, that of the physician as a Profit Maximizing Businessman. In this model, the physician will attempt to provide a set of services at a quantity which maximizes his or her net profit, given a current price and the opportunity cost of physician time, or to establish a price which will maximize income. Short-term desires of the profit maximizing physician to increase income will be tempered by long-term considerations, which require that the physician maintain an acceptable reputation and patient satisfaction in order to be able to command enough demand for his or her services in the future.

The third model of medical decision making explored in this analysis is that of the Sophisticated Physician's Target Income. In this model of physician behavior, physicians' utility from medical practice is a function of both the income that they can earn and the psychic cost of having to create demand for their own services. Therefore, a physician might seek a certain income, called the target income, but will do so only if the psychological expense of compromising personal or professional values would not be so large as to discourage the creation of demand for medical care. Therefore, in this model, induced demand is constrained by the subjective value that the physician attaches to practicing ideal medicine, either because of potential harm to the patient's utility, or because of concern about deviating from correct practice. As in the Profit Maximizing Model, induced demand will also be constrained by long-term considerations regarding reputation and detectable indicators of quality.

Since none of these three models of physician behavior has been established empirically in the health economics or medical decision making literature, we believe that it is important to consider all three in evaluating the likely impact of a volume control measure. It is likely that different physicians will be influenced by the incentives reflected in these three models to different degrees, and that, as a whole, the profession is probably influenced to some degree by all of them.

7.2. Principles of Volume Control

The goals of programs to control volume and expenditure are multiple, and are not limited simply to a reduction in expenditures for the Medicare program. Therefore, in addition to considering various models of physician behavior, our analysis has considered the various important outcomes of a volume and expenditure control program. These include a number of criteria against which a volume control can be measured. The first, control of expenditures, is the primary goal, but is constrained by other goals that are important in assuring that Medicare beneficiaries receive high quality medical care. The criteria to be discussed are the following:

1. Reduction of resource utilization
2. Reduction of expenditure
3. Quality of care
4. Access and patient equity
5. Efficiency

6. Provider equity
7. Transaction costs
8. Feasibility

1-2. Control resource utilization and expenditures. This is the primary motivation of a program to control the volume of medical services, since a reduction in volume would likely result in a reduction in expenditures. Even if there is not a reduction in volume, a reduction in the rate of increase in volume would likely result in a reduction in the rate of increase of expenditures. Realistically, it would probably be sufficient to aspire to a reduction in the rate of growth rather than an absolute reduction in Medicare outlays for Part B.

A reduction in volume, whatever it will do to Medicare's expenditures, will not necessarily reduce the use of resources (and, thus, costs) of providers proportionally. For one reason, some of the costs of health care delivery are fixed, whereas others are variable and can be eliminated with a reduction in the volume of care. A parallel phenomenon exists in Part A expenditures, in which a reduction in length of stay does not necessarily reduce the cost of hospitalization proportionally. Therefore, at least in the short-term when facilities and manpower are fixed, a reduction in the volume of services will not necessarily reduce the cost of maintaining the health care system that is currently available to provide Part B services.

Another reason that a reduction in resources consumed is not necessarily correlated with a reduction in expenditures is that changes in the way that services are classified could affect expenditures by Medicare without any effect on the services received by beneficiaries. Of course, changes in price could also affect expenditures without any necessary change in volume.

Furthermore, a short-term reduction in the volume of services currently provided need not reduce Medicare's long-term outlay for medical care services to the elderly and the disabled. While a reduction in Medicare expenditures would lead to the potential for slower premium growth for beneficiaries, reduced tax burden for the nation to support Medicare or the potential for shifting federal expenditures to other sectors of the economy (e.g., education, defense), it is also possible that reduced volume of some services currently provided by Medicare might enable the program to provide more complete coverage of services to the elderly. For example, a reduction in the expenditures on hospitalization and physician services might enable Medicare to add coverage for preventive services, pharmaceuticals, or long-term care. Therefore, a reduction in volume of a specific service should not necessarily be equated with a reduction in expenditure for the Medicare program.

3. Maintain or improve quality of care. The goal of maintaining, or even improving, the quality of care provided to Medicare beneficiaries is a major constraint on a program to control volume. To assure quality, a volume control program should be designed to avoid reduction in appropriate services and to reduce the use of harmful services. In addition, volume control should avoid undue interference with the physician's and patient's capacity to work together autonomously in seeking to optimize the patient's health and/or well-

being. Consideration of these principles of appropriateness and clinical freedom may enable a volume control program to save expenditures while not diminishing the quality of care.

- a. Avoid reduction in appropriate services: In order for a reduction in volume or in the rate of increase to occur without an excessive diminution in the quality of services provided to Medicare beneficiaries, the services which are eliminated should definitely include those with no utility (and especially those with negative utility) to the beneficiary, that is, those services which are of low quality or of low clinical value and which are unlikely to lead to improved health outcome (i.e., expected benefit is less than expected risk). However, it is also possible, indeed likely, that any feasible volume limitation will affect services which do have some value to the beneficiary which is positive, even if very small.

The difficult tasks of measuring the expected gain to the beneficiary from a medical service, and of determining the appropriate indications for services, are receiving increased attention. Until recently, this appropriateness research was generally limited to identifying when the expected clinical benefits of a service outweigh its expected clinical risk. Generally, the financial burden of the service was not considered in these investigations. However, the most effective use of Medicare funds must consider the value offered for the money spent. All would agree that a volume control program should aspire to eliminate those services with negative marginal benefit to the patient or society. More debate will be engendered regarding those services with certain small marginal positive contributions to health.

In order to carry out a classical economic analysis, marginal benefit and marginal cost must be compared. However, many health care analysts are uncomfortable measuring the benefit of medical care in the same monetary units as cost is measured, and they may, therefore, be reluctant to consider analyzing the level of services that should be provided using cost-benefit analytic methods. A popular alternative approach would be to use cost-effectiveness analysis. In this model, the health outcomes per dollar spent (the cost-effectiveness) of medical services will range across a wide spectrum. For those services for which the effectiveness is greater than zero and the net cost is greater than zero, one might define a threshold of cost-effectiveness (unit of effect per dollar of cost) below which society (as represented by Medicare) is unwilling to pay for the service, but above which society believes that Medicare should be providing the service to the Medicare beneficiary.

Therefore, any reduction in volume of services to Medicare beneficiaries should, in the ideal case, eliminate services provided to beneficiaries in which the likely outcome per unit of money spent is less than the threshold cost-effectiveness level established in some way by society as a whole. Since this threshold is difficult, if not impossible, to operationalize, it is likely that the reduction

in volume of services in response to these interventions will eliminate some services for which the cost-effectiveness is more favorable than for some services which are retained. Still, the objective of any volume control mechanism should be to exchange identification and elimination of services that are not useful to Medicare beneficiaries.

- b. Maintain as much physician and patient clinical autonomy as possible. The concept of "clinical freedom" is cherished by many clinicians, and is equated with professional autonomy. Similarly, many physicians and patients, in keeping with the model of patient agency, would suggest that physicians and patients should be free to decide upon the appropriate level of medical services without regard to external considerations such as those explored in this analysis. The issue of "clinical freedom" has been an important one in a number of Western medical care systems during the past several decades as physicians and payors of health care have explored ways of maintaining a measure of professional autonomy and clinical freedom in medical decision making, while placing constraints on the impact of that medical decision making on the public purse. An ideal mechanism for controlling the volume and expenditure of medical care would preserve clinical freedom and autonomy in clinical decision making.

4. Ensure that access and patient equity do not decline appreciably.

Access to medical care of high quality is an important central principle of the Medicare program. Therefore, any volume control system which is implemented should not only preserve that access to high quality care which currently exists, but should attempt to improve it and to ensure equity of access across all beneficiaries. Equity for beneficiaries could mean that access to appropriately high quality care does not vary with a beneficiary's spendable income or wealth. An expansion of access to high quality care is impossible without an increase in expenditures, unless there is a reduction of lower quality services that substantially reduces expenditures. This principle does not necessarily preclude individual beneficiaries from using their own resources to supplement Medicare funds. Nor does it necessarily imply completely free choice of doctor, if there are enough high quality physicians who will treat medical patients. It does suggest that any personal supplementation of Medicare payments should not limit the availability of high quality care to those beneficiaries who are unable to supplement Medicare payments.

Ensuring access to care of appropriately high quality is both a short-term and a long-term proposition. In the short-term, any volume control measure that is instituted should assure that enough physicians will be willing and able to provide high-quality care to Medicare beneficiaries at a price affordable to Medicare beneficiaries, regardless of the physician's specialty, location, or patient mix. The precise meaning of "assure," "enough," and "high quality" would need to be defined, at least implicitly. The patient's clinical problem should not influence access to care, nor should the patient's likely need for certain kinds of services make the patient more attractive to the physician than patients who need other kinds of services.

In the long-term, measures to control volume should ensure the availability of physicians in all specialties, particularly those of importance to the elderly, as well as in all locations, including rural and underserved urban areas.

5. Ensure efficiency. The principle of efficiency is incorporated into those of cost, quality and access. If Medicare beneficiaries have appropriate access to acceptable quality services at a reasonable cost, then the goal of efficiency (i.e., value for money) will be achieved. However, to attain efficiency both physicians and patients need to be sensitive to the financial implications of their decisions. At the same time as volume control measures attempt to ensure access to care and avoid economic barriers to care for Medicare beneficiaries, the same volume control measures should encourage both physicians and patients to be sensitive to the financial implications of the decisions they are making. Such an ideal volume control measure would in some way make all parties to the health care decision involved in its economic consequences, without limiting this sensitivity to beneficiaries whose financial means are meager or whose supplemental insurance is such that out-of-pocket expenditures will be required.

6. Assure equity for providers. Equity for providers is an admirable but imprecise goal. Assurance of equity generally serves the other goals described. Equity across physicians will ensure that specialty or regional groups of physicians would be rewarded similarly for their input, and that if they provide the same services, those services would be equally exposed to volume control measures. This constraint on volume control programs could help to ensure physicians' willingness or eagerness to provide care to Medicare beneficiaries and to ensure the fairness of the Medicare program to its providers.

7. Minimize transaction costs. Another criterion by which volume control measures need to be considered is the cost of implementing and administering the measure. These transaction costs provide no direct medical benefit to Medicare beneficiaries. Therefore, the transaction costs of volume control measures need to be assessed and weighed against the potential savings. These administrative or transactions costs may be incurred by the Medicare program, by the physician, or by the patient. In any case, they should be minimized.

8. Be feasible to implement. The feasibility of implementing a volume control program is an important element in its design. Many of the factors that will govern feasibility are incorporated into the goals and constraints already discussed. For example, physician participation and support by beneficiaries will require equity among physicians and restraint in interference with physicians' and patients' clinical freedom. The program's likelihood of success will depend on its ability to reduce expenditures while preserving quality and access. Its transaction costs will help determine whether the expense of implementing the program is justifiable. Other considerations with regard to feasibility include the availability of organizational mechanisms to implement the program, the capacity of information systems to support it, and the availability of data to guide decisions about the appropriateness of services.

7.3. Volume Controls

7.3.1. Guidelines

A number of parties have recently explored the opportunities for clinical guidelines to influence the volume and type of services rendered. These guidelines for medical practice may serve one of two purposes: 1) provide advice to physicians about idealized patterns of practice, or 2) enable those carrying out utilization review to define services which are inappropriate. These two uses of guidelines are sometimes described as "pathways" (idealized practice guidelines) and "boundaries" (limits for acceptable practice).

When guidelines are used for educational purposes, physicians are encouraged to practice in an idealized fashion. Educational guidelines are provided. These guidelines are based on scientific information demonstrating clear efficacy and effectiveness of the medical service being considered. In this case, the burden of proof lies with those advocating the service, but the consequences of failing to offer proof may not be serious. In contrast, when guidelines are used to provide standards for utilization review and potential denial of payment for unnecessary services, a flexible boundary of indications must be used, and the consequences of failure to remain within this boundary will be severe.

While physicians may object to guidelines as "cookbook medicine", when they are offered as educational guidelines physicians are not obligated to follow the "recipe". Educational guidelines have long been offered to physicians and many medical publishers have provided handbooks with guidelines for students and residents. Recent identification by health services researchers of wide variations in practice patterns of mature clinicians suggests that guidelines written by experts may be useful even for experienced clinicians, especially for new services or those with unclear or changing indications.

Guidelines may be designed to reduce the volume of medical care, but could cause an increase in volume if some physicians' current practices are underutilizing services, and if guidelines suggest utilization at higher levels (e.g., preventive services). Nonetheless, evidence suggests that physicians, as patients' advocates, facing financial incentives and practicing in a climate of uncertainty, do provide substantial amounts of marginal or unnecessary services. Thus, guidelines do appear to have a role in reducing unnecessary services.

The use of guidelines does have certain potential pitfalls. Guidelines may be costly to develop, and it is only feasible to develop guidelines for a limited number of services. They risk codifying current standards of practice and entrenching standards of medical care regardless of changes in technology, cost, knowledge or patients' utilities. The use of guidelines also jeopardizes the ability of physicians and patients to use their own judgment, to apply patient-specific likelihood of risk and benefit, and to incorporate the patients' personal values and risk-preferences.

Recommendations: Guidelines

Although guidelines in and of themselves would not be a strong volume control, we think they are important adjuncts to utilization review, expenditure caps, capitation and copayment. Thus, should any of these controls be implemented, we would strongly recommend the development and concurrent use of guidelines. The following points should be considered when developing or using guidelines:

- o Recent experience suggests that guidelines should be professionally derived, clearly defined, based upon data, and should demonstrate the possibility for reduction in cost while preserving or enhancing the quality of medical care or enhancing the quality of medical care at acceptable cost. Although it was felt in the 1970s that standards established by PSRO's should be local in nature, the wide variation in medical practices and the technical expertise necessary to develop guidelines has lead many experts to now believe that a set of standards should be established nationally with some room for local modification.
- o Guidelines are most likely to be effective if they are accompanied by incentives to physicians to adhere to the guidelines, or by sanctions for deviations.
- o The scientific basis for guidelines should be the same regardless of whether they are used for educational purposes (that is, as recommendations for idealized practice, or "pathways") or utilization review (that is, as criteria for negative feedback or penalties to the physicians, or "boundaries"). The design of guidelines, however, depends on their intended purpose. Specifically, guidelines developed for use in utilization review require stronger evidence before restricting reimbursement or imposing penalties on medical practices. The basis for guidelines (efficacy, safety, cost, available budget) should be specified, and can be different in different situations.
- o Guidelines should be widely disseminated, including to patients, who can then use them to evaluate the options available and to assess the care being provided to them. Guidelines should accommodate informed differences in patient preferences for risk, amenity, or convenience of access.
- o Major public-sector initiatives are needed to develop improved methods of producing guidelines, to collect better data regarding the outcomes, risk and cost of medical care, and to facilitate or sponsor the development of sets of guidelines for services that are suspected of under- or overutilization. Funding for a national center or institute should be given serious consideration.
- o Evaluation of the effectiveness of guidelines in controlling volume and/or promoting appropriate use of medical services should be promoted and sponsored.

- o Data on outcomes, risk and cost should be required for new services when Medicare decides to reimburse for their provision, and guidelines suggesting appropriate use of the services should be made available.

7.3.2. Utilization Review

A second mechanism of volume control is the implementation of utilization review systems. The prior discussion of the development of guidelines explored the criteria that might be used for utilization review and educational guidelines. In utilization review, some flexibility should be left for individual clinical judgment, for consideration of mitigating circumstances and co-existing diseases, and for patient and physician autonomy in medical decision making. Therefore, guidelines used for utilization review and denial of payment must be broad, and the burden of proof should be on those who choose to argue that the service was unnecessary.

The development of a utilization review program requires other considerations as well. These principally involve the implementation of the review, audit, and penalty processes. It appears that utilization review programs are most successful if physicians are influenced by their peers and by professional leaders and if physicians are involved in the standard setting process. Personal feedback seems to be the most effective, although there is probably a substantial "sentinel effect" due simply to the presence of the utilization review program and physicians' knowledge that their behavior is being monitored and that penalties may be forthcoming. Evidence suggests that utilization review programs do seem to have some effect, albeit small, and that those effects are likely to be "one-shot." This means that the slope of the increase in medical care costs may not be affected, although the intercept may be decreased as the curve is shifted downward with the one-time effect of a utilization review program.

Recommendations: Utilization Review

Cost-effective utilization review should be an element of any broader scheme of volume control, including the methods described in this report. In developing and implementing a UR system, the following should be considered:

- o Physicians should be involved in the process of utilization review and vested in its outcome, including participation in the design of UR criteria, involvement in its implementation, responsibility for appropriate appeals, and being influenced in a meaningful way by its findings (including the possibility of financial penalties).
- o Standards for UR should be unambiguous.
- o UR should be linked with assessment of the quality of care and designed to enhance quality.

- o UR should first be targeted to a few problems, especially those suspected to have large amounts of underuse or overuse.
- o Patients should be involved in the process and, ideally, should be affected financially by its results in order to gain their participation and interest. If they desire to override UR decisions, they should pay the cost.
- o UR should be accompanied by price adjustments to meet likely changes in demand for services resulting from the review process.
- o It is premature to initiate a full-blown UR program for Part B of Medicare. Criteria are needed; effective methods of influencing physicians are needed; and prospective review procedures should be developed.
- o It is important to involve peers and leaders of the medical community.
- o UR programs should provide personal and personalized feedback to physicians.

7.3.3. Co-payments and deductibles

The design of Medicare in 1965 included a 20% co-payment for Part B services. Since then, the desired effect of co-payment (that is, the financial involvement of beneficiaries in the cost of the care that they receive), has been eroded by the purchase of Medigap insurance policies by more than 70% of beneficiaries, and the coverage of copayment by Medicaid for another 15%. Thus, for nearly 85% of Medicare beneficiaries there now are few, if any, financial disincentives to the utilization of Medicare services.

In addition to their potential for generating an increase in Medicare expenditures, these Medigap policies are often purchased with tax subsidies. Medigap policies are subsidized by the government in two ways. First, Medigap purchasers incur much higher Medicare costs but pay the same premium as those who do not have Medigap coverage. It has been estimated that beneficiaries who purchase Medigap policies incur Medicare costs that are, on average, 39 percent higher than those who do not have such policies. In any other insurance system, higher costs are reflected in higher premiums, but in the Medicare program is not the case. Second, some Medicare beneficiaries receive Medigap coverage as a post-retirement benefit that is not considered taxable income, while others pay for coverage out of their own income, at least part of which is taxed.

Reinstating meaningful copayments, by increasing Part B premiums for those beneficiaries who purchase Medigap insurance and taxing the Medigap premiums provided as an employee benefit, is thus one approach to volume control. Such taxation would effectively increase the cost of Medigap coverage, which would in turn decrease its purchase by beneficiaries and cause them to be more concerned about the cost of services they demand.

Recommendations: Co-Payment/Deductibles

We believe that revitalizing the effect of the Part B copayment by discouraging the use of Medigap insurance is a good way to control volume and expenditures. This approach would be particularly appropriate when accompanied by a strong utilization review program based on clinical guidelines, to monitor for underutilization of services. Revitalization of copayments could be accomplished in the following ways:

- o Tax the value of employer-paid Medigap coverage as a part of retirees' income. This taxation might be targeted at Medigap policies to protect "appropriate coverage," for example for services thought to be underutilized, where underconsumption is feared, or for services that are not covered by Medicare.
- o Institute a surcharge on Medigap policies, for example added to the Part B premium and related to the beneficiaries' income or wealth.

7.3.4. Capitation

Although health services investigators have demonstrated that health maintenance organizations are capable of reducing expenditures, principally through a reduction in hospitalization and length of stay, HMOs have not been shown to moderate the rate of increase of health expenditures over time. The likelihood of capitation having a substantial impact on Medicare expenditures will depend in part upon the risk arrangement that is implemented and the organization of the capitation programs.

The reason for capitation having a salutary effect on volume is not clear. While some observers believe that physician incentives are important, others have suggested that the group process of physicians working together and the effect of capitation on the organization of practices are the principal influences. Therefore, it is possible that similar changes in practice organization could lead to different practice styles independent of a change in financing mechanisms. Case management may also play a role, though primarily in conjunction with financial controls.

Although capitation has the potential to control Part B volume and expenditures, a number of problems related to capitation make it less attractive as a volume control than other approaches. Organizational changes necessary for a completely capitated system would be difficult to realize in the short term (and maybe even in the longer term). Capitation on a partial basis raises the possibility of increased referrals for services not included in the partial capitation, which could preclude significant cost savings. Either partial or complete capitation present risk pooling problems that require case mix severity adjustments; at present, these methods are not well developed. Likewise, either type of capitation requires an equitable determination of capitation fees. Given the current problems with AAPCC's for Medicare HMOs, a successful approach to setting appropriate and equitable fees seems unlikely in the near future.

Recommendations: Capitation

The administrative and design problems associated with capitation of Part B only make it a less attractive option (at this point in time) than expenditure caps with more rigorous utilization review or adjustments in copayment. To make capitation more feasible in Part B, the following should be undertaken:

- o Develop techniques to measure severity of disease or other predictors of medical care utilization in order to establish prices for capitated services.
- o Initiate and evaluate selected programs of Partial Physician Capitation, with groups of physicians chosen for capitation of physician services only on the basis of their site of care or the type of illness being treated.
- o Encourage case management techniques, which might be utilized in fee-for-service practice (especially in conjunction with utilization review) as well as in capitated practice.

7.3.5. Expenditure Growth Targets

In the face of recent reductions and freezes in prices paid by Medicare for physician services, some have concluded that an explicit process of defining the target for Medicare expenditure growth would provide more predictability and control for both Medicare and the physician community. Expenditure targets for future time periods would be derived by determining the appropriate growth in aggregate expenditures for a given geographical region over a given period of time. Thus, expenditure targets would be defined based upon a population of patients, not a population of physicians. If expenditures were to exceed this target, then various mechanisms to change Medicare fee levels could be set in place to set total Medicare expenditures right, either in the current or in the next time period. Geographic regions could be established based upon the uniformity of input prices, variation in medical practice, the degree to which the physician community is likely to be able to exert influence, and a number of other characteristics.

All these models would predict that an expenditure target would enable Medicare to assure beneficiaries that their premiums and the outlays on behalf of their medical care would increase in a predictable fashion (presuming limits on balance billing). Expenditure growth targets cannot fail to control the growth in expenditures. They also would enable the medical profession, in keeping with the Patient Agency Model, to take responsibility for the control of volume and services and would forge new professional relationships and possibly new physician organizations. For example, with this volume control, peer interaction and review or case management programs, such as exist in many prepaid practices, might be instituted.

It would be desirable for the prices paid for medical services to be considered satisfactory by physicians before an expenditure target system was

implemented, in order that physicians be willing or even eager to participate in such a system. Because there are, at present, perceptions of inequity in reimbursement among physician specialties (many of which would be addressed with the RBRVS), a system that makes one physician financially liable for the practice patterns of other physicians would likely meet significant resistance.

While expenditure targets could be mandatory, they could also include certain elements of voluntarism. Within a mandatory expenditure target system, groups of physicians could elect to "opt out" and to practice within subsystems of the expenditure target population. For example, health maintenance organizations or preferred provider organizations that have demonstrated their ability to provide care efficiently might choose to be providers with targets for their present populations in order to free them from the potentially higher utilization rates of the general physician and patient populations.

A number of design issues exist with regard to expenditure targets, including the method of projecting expenditures, the method of sharing overrun costs or savings, the timing of adjustments, the size of the geographic area, the degree to which assignment would need to be mandatory, and the issue of mandatory vs. voluntary participation. A number of concerns also exist, including the perceived fairness for physicians who are held accountable financially for the decisions of their peers, the potential that individual physicians would not limit their own utilization of services and that a spiral of increase in volume and decrease in price would occur.

Recommendations: Expenditure Targets

We believe that expenditure targets, especially in conjunction with rigorous utilization review based on sound clinical guidelines, offer an attractive approach to control of Part B expenditures. Expenditure targets are sure to control expenditures, and they provide Congress, HCFA and beneficiaries a predictability about Medicare expenses and premiums that is presently lacking. They also allow for a mechanism by which costs associated with changes in technology or in the health status of the Medicare population can be dealt with. We suggest that the following factors should be considered in the implementation of expenditure targets:

- o Appropriate populations of Medicare beneficiaries must be selected for inclusion. The following considerations enter into the decision about the geographic area covered by a target:
 1. The population covered should be large enough to be administratively feasible and to avoid a large number of small administrative units.
 2. The physician population within a target area should be small enough to allow for peer interaction and influence.

3. Current organizational structures could be used to administer the target, but new organizations should also be considered.
 4. The size of the population should be large and stable enough to avoid large year-to-year fluctuations in expenditures, in order to make the targets predictable.
- o Symmetric incentives should probably be offered (i.e., decrease in unit prices for exceeding the target and increase for meeting it), at least initially, but neither increases nor decreases in price need be equal to the difference between the target and actual expenditures.
 - o To increase acceptance of targets by the physician population an appropriate fee schedule should be developed before implementation of an expenditure target.
 - o Part A expenditures should be considered in setting the expenditure target since some shifting of costs from Part B to Part A may occur.
 - o Mandatory assignment is not necessary but some limit on out-of-pocket expenditures for beneficiaries is desirable.
 - o A pilot program with voluntary participation should be instituted to assess the administrative and quality issues raised.

7.3.6. Collapsed Procedure Codes

Given the large number of codes for medical services, it is possible that physicians may choose to "upcode" by indicating that the service they provided was one with a slightly more complicated level of severity or intensity. Thus, collapsing procedures into fewer codes could potentially reduce the opportunity for upcoding. However, our analysis suggests that the net effect of collapsing codes is unpredictable. In essence, collapsing codes in balanced-budget fashion reduces the incentive to upcode by increasing the "distance" between two codes, but this may be offset by the increased incentive to "jump the gap" due to the higher price differential between two codes.

Recommendations: Coding

Because the net effect on volume and expenditures of collapsing procedure codes is unpredictable, we do not recommend it as a Part B volume control. If collapsed coding is considered for implementation at some point in the future, we recommend the following prior to implementation:

- o Patients should be involved in verifying the accuracy of coding.
- o Experimentation is needed with new techniques for coding visits.

7.3.7. Bundling of Procedures

It has been argued that the presence of more than 7,000 codes for medical services induces physicians to disaggregate the services that they provide into more than one service in order to bill for each service separately when separate codes for different services exist. Thus, bundling of services into logical clinical packages might be desirable because it would improve physician incentives to control volume, might be a step on the road to full capitation, and could reduce the opportunities for "gaming" by physicians. While the option of bundling does have promise for decreasing the ability of physicians to disaggregate their services and to increase the number of bills rendered for the same service, bundling also presents the risk that physicians would bill for a bundled set of services and yet underserve the patient by providing only some of those services which were originally understood to be in the package. It is also possible that physicians would substitute services outside the package to avoid the limit on payment.

Recommendations: Bundling

Bundling should be approached with caution since it could result in distorted incentives for provision of services not included in the bundle. At present, we do not recommend bundling as a volume control. Should it be considered in the future, these factors should be incorporated into its design:

- o Since the idea behind bundling is to reduce the incentive for physician-induced demand, the greater the prospects for demand creation, the more useful it will be to bundle.
- o Services typically provided by the principal physician in a fee-for-service system should be included in the bundle.
- o The more services are substitutable between those performed by the principal physician and those performed by other providers, the more appropriate it will be to include them in the bundle (so as to provide an incentive for the primary physician to provide them in the lowest-cost manner).
- o Conversely, if a service is strongly complementary to the principal service and therefore likely to be provided regardless of whether it is included in the bundle, it need not be included in the bundle since doing so will increase the amount of cost for which the doctor is put at risk.
- o The less variation in severity of disease and in practice patterns that accompany a primary service, the more appropriate it will be to bundle that service with others clinically associated with it.

7.4. Surveys of Commercial Insurance Firms and Medical Carriers

All Medicare and most private sector health insurance carriers have utilization review programs, although Medicare carriers appear to focus more on reviewing the appropriateness of physician services than do private carriers. This is partly due to the presence of mandated prepayment screens for Medicare. However, many Medicare carriers had these screens before they were mandated and, in some cases, their pre-mandate screening parameters were tighter than those currently used.

Several mandated prepayment screens were cited frequently as ineffective, whereas other optional screens were suggested for inclusion in the list of mandated screens. A systematic study of the effectiveness of prepayment screening is warranted, and such a study is currently underway at the HCFA Research Center.

Medicare carriers suggested that new screens be introduced along with programs to educate physicians about these screens. In general, implementation issues are important and should not be ignored. On the other hand, provider resistance to prepayment screens is not necessarily a sign that they are ineffective.

Medicare carriers operate vigorous postpayment review programs. In many cases, physicians are reviewed if their practice patterns are more than 2 standard deviations from the norm. However, the criteria for postpayment review are not uniform. Standards based on reviewing a prespecified number of physicians may be difficult to justify in terms of detecting unusual practice patterns.

Private insurance carriers were found to operate inpatient utilization review programs that combine pre-admission certification, concurrent review, and retrospective review of hospital inpatient care. The respondents estimate that inpatient utilization review reduces total cost by about 7 percent. Many believe that pre-admission certification increases cost and utilization outside the hospital, however.

For managed fee-for-service programs, utilization review has been largely directed at reducing the use of inappropriate inpatient services. PPOs have made far less effort in selecting preferred physicians than hospitals, often using hospital staff privileges as the major screening criterion. HMOs have historically realized their savings by reducing the number of hospital days.

Private carriers have been slow to innovate in the area of physician payment. All but 8 of the firms surveyed by HIAA still pay usual and customary charges. Only 6 respondents utilized "controlled" methods of paying physicians. Commercial insurers appear to place more emphasis on physician payment reform in their PPO arrangements, which frequently use discounted charges or fee schedules to pay physicians. The average discount was estimated to be about 14 percent. This was exceeded in importance by the cost-saving effect of utilization review.

Perhaps the most significant result from our surveys is that neither public nor private carriers have taken an innovative approach toward bundling physician services and collapsing the codes used for paying physicians. The private respondents, by and large, were ignorant of this concept; Medicare carriers generally thought that they had to recognize HCPCS codes for billing. Medicare carriers appear to be ahead of the private sector in using global fees for surgery, but otherwise they have not attempted to bundle physician services into broader reimbursement packages.

Physician education and feedback is utilized more extensively by the Medicare carriers than by private insurers. This may be due to the fact that they lead the private sector in terms of postpayment review, and this is the area where physician education programs are most likely to occur. Medicare carriers, in general, are sensitive to the need for physician education and information programs.

7.5. Impact of Resource Based Relative Value Scale

Under a revenue-neutral implementation of RBRVS with mandatory assignment, the price of physicians' services will increase for some services and decline for others, but the change (if any) in the total cost and volume of physicians' services is impossible to predict with certainty. All scenarios rest on untested behavioral assumptions, including the willingness of patients to accept the newly more profitable services and the appropriate model of physician behavior to use to predict physician response.

The most important source of ambiguity comes about because a price change has two conflicting theoretical effects on physician incentives. On the one hand, a physician will want to use fewer of the less profitable services and more of the more profitable ones to treat a condition -- a substitution effect. On the other hand, if the service whose price is cut is an important part of the physician's total business, the price cut will cause income to fall unless he or she can increase demand for other services. If income falls, the physician may create demand for the new lower-priced service to get income back closer to a target level -- an income effect. It is not possible to determine a priori which effect will predominate, and so it is not possible to determine whether aggregate volume will rise or fall with a resource based relative value scale.

REFERENCES

- Brook RH, Williams KN, Rolph JE: Use, costs, and quality of medical services: Impact of the New Mexico peer review system. *Annals of Internal Medicine*, 1978; 89:256-263.
- Buck CR, White KL: Peer review: Impact of a system based on billing claims. *ENJM*, 1974; 291(17):877-883.
- Buczko, William: Physician utilization and expenditures in a Medicaid population. *Health Care Financing Review*, 1986; 8(2):17-26.
- Burney I, Hickman P, Paradise J., and Schieber G: Medicare physician payment, participation and reform. *Health Affairs*, 1984; Winter:5-24.
- Blumberg MS: Health status and health care use by type of private health care coverage. *Milbank Memorial Fund Quarterly*, 1980; 58:633.
- Chassin MR, Brook RH, Park RE, Keesey J, Fink A, Kosecoff J, Kahn K, Merrick N, and Solomon DH: Variations in the use of medical and surgical services by the Medicare population. *NEJM*, 1987; 314(5):285-90.
- Connell FA, Blide LA, Hanken MA: Clinical correlates of small area variations in population-based admission rates for diabetes. *Med Care*, 1984; 22(10):939-49.
- Danzon PM, Manning WG Jr., Marquis MS: Factors affecting laboratory test use and price. *Health Care Financing Review*, 1984; 5:23-32.
- Danzon PM: Economic factors in the use of laboratory tests by office-based physicians. Publication R 2525-1-HCFA. Santa Monica, CA, The Rand Corp. 1980.
- Danzon PM, Manning WG, Marquis MS: Factors affecting laboratory test use and price. Publication R-2897-HCFA. Santa Monica, CA, The Rand Corp. 1983.
- Dranove D: An economic model of the physician-patient relationship, dissertation. Stanford (CA) University, 1983.
- Dyck FJ, Murphy FA, Murphy JK, et al: Effect of surveillance on the number of hysterectomies in the province of Saskatchewan. *NEJM*, 1977; 296:1326-8.
- Eisenberg, JM, Myers LP, Pauly MV: How will changes in physicians payment by Medicare influence laboratory testing? *JAMA*, 1987; 258(6):803-08.
- Eisenberg JM: Doctor's Decisions and the cost of medical care. Ann Arbor, Mich, Health Administration Press, 1986.
- Eisenberg JM: Physician utilization the state of research about physicians' practice patterns. *Medical Care*, 1985; 23(5):461-83.

- Feldstein PJ, Wickizer TM, Wheeler JRC. Private cost containment. The effects of utilization review programs on health care use and expenditures. NEJM 1988, 318(20):1310-14.
- Feldstein PJ, Wickizer TM, Wheeler JR: Private cost containment. NEJM 1988; 318(20):1310-1314.
- Fisher, CR: Impact of the prospective payment system physician charges under Medicare. Health Care Financing Review, 1987; 8(4):101-103.
- Foxman B, Valdes RB, Lohr KN, Goldberg GA, Newhouse JP, Brook RH: The effect of cost sharing on the use of antibiotics in ambulatory care: Results from a population-based randomized controlled trial. J. Chronic Dis 1987; 40(5): 429-37.
- Gabel JA, Rice TH. Reducing public expenditures for physician services: The price of paying less. J. Health Politics, Policy and Law 1985; 9(4):595-609.
- Hammons GT, Brook RH, Newhouse JP. Changing physician payment for Medicare patients. Projected effects on the quality of care. Western J. of Medicine 1986; 145(5):704-09.
- Hammons GT, Brook RH, Newhouse JP. Selected alternatives for paying physicians under the Medicare program. Effects on the quality of care. Santa Monica: The Rand Corporation, June, 1986.
- Hillman A. Financial incentives for physicians in HMOs: Is there a conflict of interest? NEJM, 1987; 317:1743-48.
- Holahan J, and Scanlon W: Physician pricing in California: Price controls, physicians' fees, and physicians' incomes from Medicare and Medicaid. Washington DC: U.S. Dept. of Health and Human Services, Health Care Financing Administration, 1979.
- Hsaio WC, Stason WB: Toward developing a relative value scale for medical and surgical services. Health Care Financing Review, Fall 1979; 1(2):23-38.
- Hornbrook MC, Berki SE. Practice mode and payment method effects on use, costs, quality and access. Medicare Care 1985; 23(5):484-511.
- Imperiale TF, Siegal AP, Crede WB et al.: Preadmission screening of Medicare patients. JAMA, 1988; 259(23):3418-3421.
- Jencks SF, Dobson A.: Strategies for reforming Medicare's physician payments: physician diagnosis-related groups and other approaches. NEJM 1985; 312(23):1492-99.
- Juba, DA: Medicare physician fee schedules: issues and evidence from South Carolina. Health Care Financing Review, 1987; 8(3):57-67.

- Kosecoff J., Karouse DE, Rogers WH, McCloskey L., Winslow CM, Brook RH:
Effects of the National Institute of Health Consensus Development Program
on physician practice. JAMA, 1987; 258(19):2708-13.
- Leader S., Guildroy J., Kennan S., Lehrmann E., Skinner E: The Canadian
health care system: A special report on Quebec and Ontario. American
Association of Retired Persons, 1988.
- Liebowitz A, Manning WG, Newhouse JP: The demand for prescription drugs as a
function of cost-sharing. Soc Sci Med 1985; 21(10):1063-9.
- Luft HS, Hunt SS, Maerki SC: The volume-outcome relationship: Practice-
makes-perfect or selective-referral patterns? Health Services Research,
1987; 22 (2):157-182.
- Luft HS: HMO performance: Current knowledge and questions for the 1980s.
A research agenda considered. Group Health J, 1980a; 1:34.
- Luft HS: Trends in medical care costs. Do HMOs lower the rate of growth?
Medical Care, 1980b; 18(1):1-16.
- Luft HS: Health maintenance organizations: Dimensions of performance.
New York: John Wiley & Sons, 1981.
- Manning WG, Liebowitz A, Goldberg GA, Rogers WH, Newhouse JP: A controlled
trial of the effect of a prepaid group practice on use of services. NEJM
1984; 310(23):1505-10.
- Marquis MS: Laboratory test ordering by physicians: the effect of
reimbursement policies. Publication R-2901-HCFA. Santa Monica, CA, The
Rand Corp., 1982.
- McCarthy EG, Widmer OW: Effects of screening by consultants on recommended
elective surgical procedures. NEJM, 1974; 291(25):1331-1335.
- McPherson K., Wennberg JE, Hovind OB, Clifford P.: Small area variations in
the use of common surgical procedures: an international comparison of
New England, England and Norway. NEJM, 1982; 307:1310-4.
- Milstein A., Oehm M., Alpert G.: Gauging the performance of utilization
review. Business and Health, February 1987: 10-12.
- Mitchell JB, Cromwell J, Calore KA, Stason WB: Packaging physician services:
alternative approaches to Medicare Part B reimbursement. Inquiry, (W)
1987; 24:324-43.
- Mitchell JB: Physician DRGs. NEJM 1985; 313(11):670-5.
- Mitchell JB, Cromwell J, Calore KA, Stason WB: Packaging physician services:
Alternative approaches to Medicare Part B reimbursement. Inquiry 1987;
24:324-43.

- Moore SH, Martin DP, Richardson WC. Does the primary care gatekeeper control the costs of health care? Lessons from the SAFECO experience. NEJM, 1983; 309(22):1400-1404
- Newhouse JP: Has the erosion of the medical marketplace ended? J. of Health Politics, Policy and Law, 1988; 13(2):263-78.
- Newhouse JP, Schwartz WB, Williams AP, Witsberg C: Are fee-for-service costs increasing faster than HMO costs? Medicare Care, 1985; 23(8):960-66.
- Newhouse JP, Manning WG, Morris CN, Orr LL, Duan N., Keeler EB, Leibowitz A., Marquis KH, Marquis MS, Phelps CE, Brook RH: Some interim results from a controlled trial of cost sharing in health insurance. NEJM 1981; 305(25):1501-7.
- O'Grady KF, Manning WG, Newhouse JP, Brook RH: The impact of cost sharing on emergency department use. NEJM 1985; 313:484-90.
- Palmer RH: The challenges and prospects for quality assessment and assurance in ambulatory care. Inquiry 1988; 25:119-31.
- Paris M., McNamara J., Schwartz M.: Monitoring ambulatory care: Impact of a surveillance program on clinical practice patterns in New York. AJPH, 1980; 70(8):783-788.
- Pauly MV: The ethics and economics of kickbacks and fee splitting. Bell J. of Economics, Spring, 1979.
- Pauly MV and Langwell K.
- Pauly MV: Doctors and their workshops. Chicago, University of Chicago Press, 1980.
- Pauly MV and Satterwaite M: The pricing of primary care physicians' services: a test of the role of consumer information. Bell J. of Economics, Autumn, 1981, 488-506.
- Phelps, C.: "Induced demand - can we ever know its extent? Journal of Health Economics, 1986; 5(4):355-365.
- Physician Payment Review Commission: Annual Report to Congress. Washington, DC: March 1, 1988.
- Rice T., McCall N: Changes in Medicare reimbursement in Colorado: impact on physician's economic behavior. Health Care Financing Review, 1982; 3:67-85.
- Roddy PC, Wallen J, Meyers SM: Cost sharing and use of health services: The United Mine Workers of America Health Plan. Medical Care 1986; 24(9):873-78.

- Roos NP, and Roos LL: High and low surgical rates: risk factors for area residents. Am J Public Health , 1981; 71:591-600.
- Scheffler RM: The United Mine Workers' Health Plan: An analysis of the cost-sharing program. Medical Care 1984; 22(3):247-54.
- Schroeder SA, Myers LP, McPhell SJ, Showstack JA, Simborg DW, Chapman SA, Leong JK: The failure of physician education as a cost containment strategy. Report of a prospective controlled trial at a university hospital. JAMA, 1984; 252(2):225-300.
- Schulenburg JMG: Report from Germany: Current conditions and controversies in the health care system. J. Health Politics, Policy and Law 1983; 8(2):320-51.
- Schwartz, JS, Williams SV, Kitz DK and Eisenberg JS: Effectiveness of a statewide utilization review program using profile analysis. Submitted for publication to Medicare Care, 1988.
- Scitovsky AA, McCall N: Use of hospital services under two prepaid plans. Medical Care, 1980; 18-30
- Sisk JE, McMenamin P, Ruby G, Smith ES. An analysis of methods to reform medicare payment for physician services. Inquiry 1987; 24:36-47.
- Siu AL, Sonnenberg FA, Manning WG, Goldberg GA, Bloomfield ES, Newhouse JP, Brook RH: Inappropriate use of hospitals in a randomized trial of health insurance plans. NEJM 1986; 315:1259-66.
- Sorenson A, Wersinger R, Seward E et al.: Health status, medical care utilization and cost experience of prepaid group practice and fee-for-service populations. Group Health Journal 1981; 2:4
- Taylor AK, Farley Short P, Horgan CM: Medigap insurance: friend or foe in reducing Medicare deficits? In: Health Care in America The Political Economy of Hospitals and Health Insurance, HE Frech III, Ed. San Francisco: Pacific Research Institute for Public Policy, 1988.
- U.S. Congress, Office of Technology Assessment: Payment for Physician Services: Strategies for Medicare, OTA-H-294 (Washington, DC: U.S. Government Printing Office, February 1986).
- Wennberg JE and Fowler FJ Jr.: A test of consumer contribution to small area variations in health care delivery. J Maine Med Assoc., 1977; 68:275-9.
- Wennberg J. and Gittelsohn A.: Variations in medical care among small areas. Sci Am, 1982; 264(4):120-35.
- Wennberg J: Dealing with medical practice variations: A proposal for action. Health Affairs 1984; 3(2):38-45.

Wennberg J.: Commentary on patient need, equity, supplier-induced demand, and the need to assess the outcome of common medical practices. Medicare Care, 1985; 23(5):512-21.

Wilensley GR and Rossiter LF: The relative importance of physician-induced demand in the demand for medical care. Milbank Mem Fund Q 1983; 61:252-77.

Yett DE, Der W, Ernst RL, et al. Physician pricing and health insurance reimbursement. Health Care Financing Review 1983; 5(2):69-80.

Appendix I: Financial Incentives in
Medical Decision Making

FINANCIAL INCENTIVES IN MEDICAL DECISION MAKING

Eugene Rich, M.D.

12/1/88

Background

Physicians and policy makers are directing renewed attention to the role of financial incentives in medical decision making (Roper, 1988). Increasingly, the mechanism of physician payment is recognized as a potential source of distortion for clinical practice. Current fee-for-service models are viewed as over-stimulating the use of technology (PPRC, 1987), penalizing the provision of "cognitive" and preventive services (Berenson, 1987), and contributing to the ongoing shifts in specialty choice by new physicians (McCarty, 1988). Alternative reimbursement schemes such as capitation are viewed as no panacea; many physicians express concerns that financial and administrative pressures in prepaid health plans may also adversely affect their practice (Hillman, 1988; Levinson, 1987).

These observations are disturbing when juxtaposed with the traditional model of the physician as the patient's agent. It is one thing for a health care professional to complain about what he or she is paid; it is quite another to imply that the mode of payment alters what the professional will do. Physicians may appear to be failing their responsibility to patients if personal incentives systematically influence their recommendations. A

correct understanding of the magnitude and mechanism of such influences has important implications for the financing and organization of medical practice. If corruption, ignorance and venality are the predominant culprits, then practice distortions due to financial incentives should be addressed by heightened professional credentialing, peer review, and professional censorship. If instead, financial incentives have more subtle and pervasive effects, then more sophisticated solutions will be indicated.

Decision Rules: A Model For Medical Decision Making

We postulate that the process of medical problem solving evokes fundamental limitations in human decision makers. These limitations allow for influence by financial and other personal incentives, even while physicians attempt to serve the patients' interest. To illustrate this we will explore the implications of financial incentives on physicians using a simplified model of medical decision making. We will assume that to solve problems, physicians must draw upon a vast array of remembered decision rules. These rules may be formulated as: "for patients with condition A, take action X" (Eddy, 1982). Clinical decision making can then be modeled as a series of individual decisions. First the physician must learn a decision rule, such as "if A, then X" (Figure I). Having learned this decision rule, the physician must recognize cases of A (Figure II). Finally, the physician, having decided that the patient represents a case of

A, and requires intervention X, must undertake to have the intervention performed for the patient (Figure III). To this simplified model numerous additional steps could be added, such as identifying the universe of decision rules, remembering the learned rule, remembering A and X when the patient also has condition D requiring action Z, developing of a clinic system to facilitate X, etc. For simplicity, our model will focus on these three elements of decision making, which should be sufficient to illustrate the role of financial incentives. Using our model, we will review the evidence for systematic effects of financial incentives on such decisions and will consider some of the potential consequences of such effects.

Learning Decision Rules

The first element of this model of medical decision making is learning decision rules (Figure I). Such heuristics can have a wide variety of sources, varying in both formality and authority (Eddy, 1982; Rich, 1985). Certainly some physician learning comes from CME courses and text books (Curry and Putnam, 1981), but much comes from the less formal sources of colleagues and consultants (Covell, 1985; Greer, 1988). If medical opinion were unified on the appropriate action to take for any specified condition, then variation of medical practice might be predominantly due to ignorance. Since for so many clinical practices, scientific

FIGURE I
LEARNING

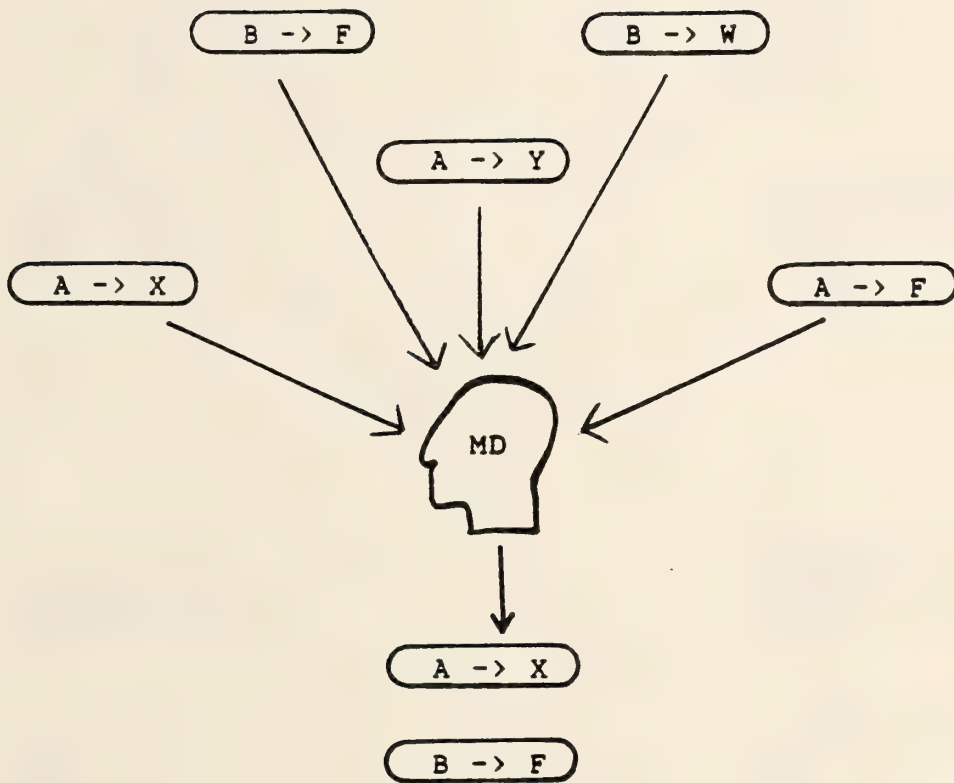


FIGURE II
DIAGNOSIS

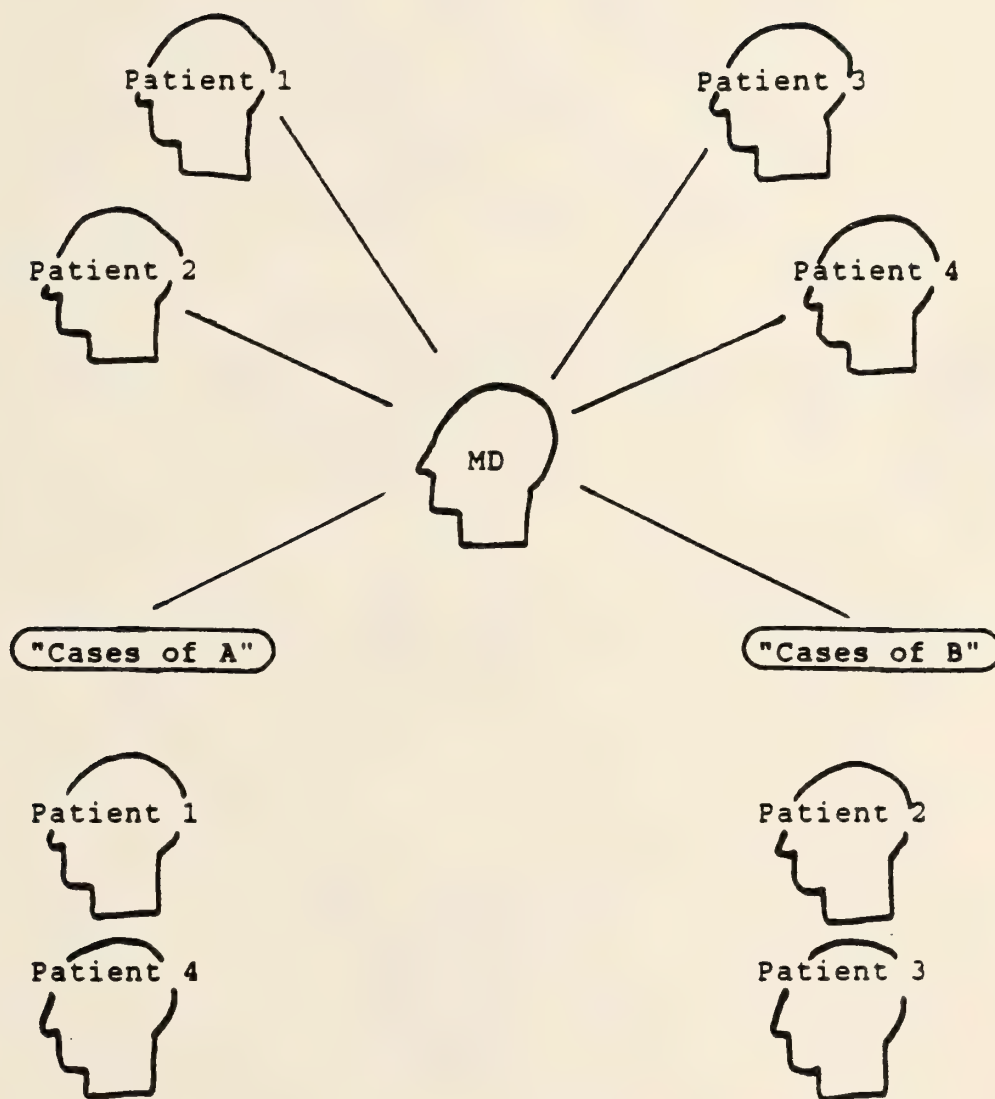
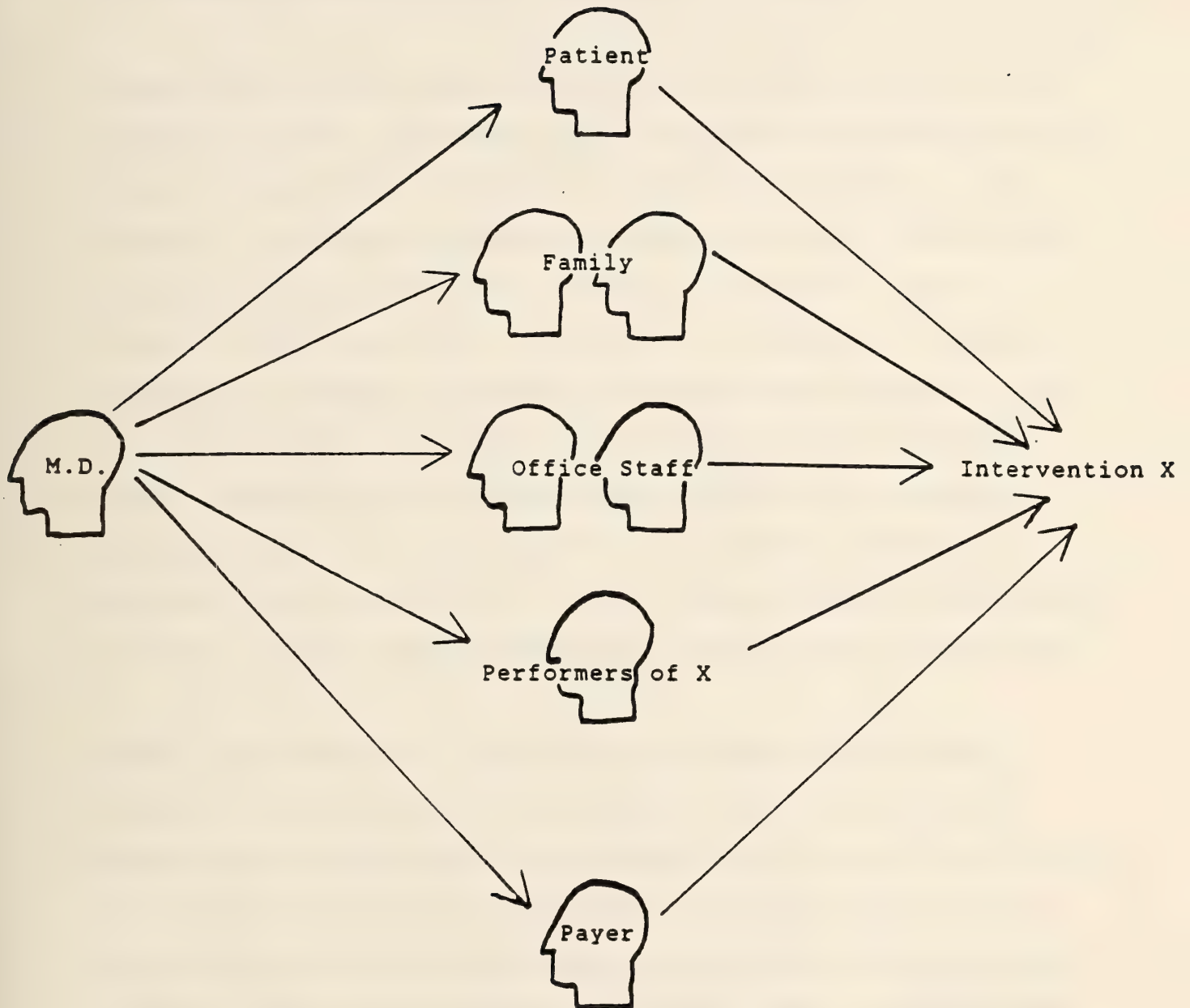


FIGURE III
INTERVENTION



evidence is absent or inadequate, and authoritative opinions are imprecise or contradictory, physicians are faced with a variety of potential decision rules from which to choose.

Outside the medical arena, beliefs not based on certainty are termed "opinions" (Websters, 1979), and various research suggests these to be subject to personnel financial incentives. For example, in studies of gambling, estimates of the likelihood of events are influenced by the monetary value of the events (Slovic, 1966; Lee, 1971). Similarly, in studies of cognitive dissonance, subjects' opinions have been shown to shift in the direction favored by reimbursement (Festinger and Carlsmith, 1959). Payment has also been shown to alter the strategies decision-makers use when interpreting evidence (Waller and Mitchell, 1984; Arkes, et al, 1986). Thus, financial incentives clearly influence opinions when the correct answer is ambiguous.

There is evidence that financial incentives may influence opinions in medicine as well. Hartley found that American physicians in fee-for-service medicine viewed renal complications of hypertension as more likely than did British physicians in similar practices (and the American physicians performed renal function tests more frequently) (Hartley, 1987). Hlatky found that physicians in fee for service practice viewed certain types of standardized patients as requiring angiography; the identical standardized patients were not so viewed by physicians in capitated practice (Hlatky, 1983). In a randomized controlled

trial, Hickson demonstrated that fee-for-service reimbursement influenced physician decision making for scheduling discretionary return appointments (Hickson, 1987). Thus evidence from psychological research and from medical practice is consistent with the thesis that when the correct answer is unclear, financial incentives influence the selection of competing decision rules.

Recognizing Cases

The next step in our model is the recognition of cases. Even if a physician accepts the decision rule, "if A then X", X's will only occur when the physician "diagnoses" cases of A (Figure II). It should be noted that a case need not be a "disease", but could be a risk factor (if smoker, prescribe Nicotine gum) a symptom (if chest pain, order electrocardiogram) or a combination of traditional "diseases" (if acute pulmonary thromboembolism and acute upper gastrointestinal hemorrhage then place vena cava umbrella). In each case, however, the physician recognizes condition A and thereupon undertakes action X.

If physicians were always successful in identifying patients with these conditions from among all the patients they see, then variation could only occur from ignorance or neglect. Since the process of recognizing cases of A is inherently subject to error, the probabilistic concepts of diagnostic accuracy apply (Swets, 1988). The physician's sensitivity to A is the chance of the

physician recognizing A, when confronted with patients having A. Specificity is the chance of correctly recognizing that A is absent when seeing a patient who does not have A. For any diagnostic modality with a fixed diagnostic accuracy (whether human or machine), sensitivity can only be enhanced at the cost of specificity. With human observers, it has long been recognized that financial incentives for "diagnosis" enhance sensitivity at the loss of specificity (Green and Swets, 1966).

Within the medical arena, physicians have been shown to over-diagnose some conditions (Dawson, 1987) and under-diagnose others (Korn, 1988). Several studies of medical practice have found systematic differences suggesting variations in diagnosis coincident with financial incentives to the physician (Haynes De Regt 1986; Francis 1984).

If the findings of financial incentives in signal detection do indeed hold for medical diagnosis, then incentives favoring the recognition of "condition A" would result in increased sensitivity and decreased specificity. Under most existing "fee-for-service" reimbursement schemes physicians are not usually reimbursed for diagnoses themselves. Financial incentives would also apply when action X (which is linked to condition A in the decision rule "if A then X") is advantageously reimbursed. Financial incentives which could enhance specificity over sensitivity include financial disincentives for performing action X (e.g. the perceived physician cost of X exceeded revenue) and

penalties assessed for erroneously performing X (e.g. risk of law suit or PRO censorship when X is done on false A's).

Performing Interventions

Once the physician has accepted the rule "if A then X" and has recognized the patient as an example of condition A, then action X should result. In clinical medicine, however, a series of additional events may need to occur, each of which introduces possible variation. For X to actually occur, the physician must interact successfully with the patient, clinic personnel, the payor organization and others (Figure III). Research outside the medical field has demonstrated that financial incentives facilitate the completion of various tasks. Not surprisingly, physicians seem able to improve patient compliance for financial gains (Hickson, 1987).

Thus, in this decision rule model of medical decision making, practice variations susceptible to financial incentives may occur at the level of choosing decision rules, recognizing cases, and performing interventions. Throughout this process the physician is applying a decision making strategy to serve the patient's interest. In using this strategy, the physician must resolve such relevant ambiguity as the quality of the medical evidence (choosing rules), the clinical findings of the case (recognizing cases), and the preferences of the patient (performing interventions). The intrusion of financial incentives into this

process may be inescapable in the face of such ambiguities (e.g. gambling behavior, cognitive dissonance, signal detection). Thus, the physician decision maker would perceive him or herself as the patient's agent, often unaware of the influence of personal incentives. Nonetheless the resulting decisions may be substantially confounded by these incentives.

Selective Incentives for X vs. Y

We will next consider a series of scenarios which illustrate the practical implications of such financial incentives on physician decisions. In these illustrations two different decision rules are viewed as competing for acceptance by the physicians ("if A then X" versus "if A then Y"). These competing decision rules are treated as equally well supported by scientific evidence and/or authoritative opinion. Scenario I illustrates the expected frequency of application of X and Y in a population of 100 physicians and 20,000 patients (10,000 of whom actually have condition A), where the physicians have a sensitivity of 90% and specificity of 90% for condition A, and where patient compliance with treatment is ordinarily 90%. Figure IA demonstrates the expected rates of X and Y where there are no selective financial incentives for one treatment over the other. This circumstance would obtain, for example, where the ratios of perceived revenues to expenses were similar for each of each possible actions "X", "Y", and "no treatment".

Figure IA starts with a population of 100 physicians who must choose either "if A then X" or "if A then Y". Since these decision rules are both known to all 100 physicians and are equally well supported by objective evidence, the physician belief rate for each rule is 0.5. Thus 50 physicians choose "if A then X" (50 X MD) and 50 choose "if A then Y" (50 Y MD). Each physician now sees 200 patients in practice. One hundred of these patients actually have condition A (prevalence of A 0.5). The other 100 have condition B, which requires "no treatment". The physicians are assumed to have a sensitivity for A of 0.9 and a specificity also of 0.9. Each of the 50 "X MD's" therefore correctly recognize 90 cases of A and 90 cases of B, falsely diagnosing 10 cases of A and 10 cases of B. As a result the X MD's recommend their intervention to a total of 5,000 patients (4,500 with A and 500 falsely diagnosed as A), the Y MD's do likewise. In each group 90% of the patients actually obtain the recommended intervention (compliance rate 0.9). Thus, of the 20,000 patients seen by the 100 physicians, 4,500 receive intervention X (of whom 450 didn't need intervention), 4,500 receive Y (again 450 unnecessarily), and 900 do not receive the needed intervention, either because of failure to diagnose or failure in "compliance".

In Figure IB we illustrate in succession the three levels of potential confounding induced by a selective financial incentive for X. We assume that this incentive for X induces a change in the physician's likelihood of accepting the decision rule X

(physician belief rate from 50 to 55%), a change in the physician's likelihood of diagnosing A (shift in sensitivity from 90 to 95% and shift in specificity from 90 to 85%), and a change in the physician's likelihood of accomplishing intervention X (a change in the compliance rate from 90 to 95%). Figure IB demonstrates that even such modest financial incentives could result potentially in substantial differences in the application of one technology over another, even when the evidence for the benefit of these two technologies is balanced. As expected, each additional source of confounding increases the proportion of treated patients who receive the financially advantaged treatment (X) from 50% in the no incentive case to 55% where the influence is only at the level of choosing the rule, to 57% where there is influence both on choosing rules and recognizing cases, to 59% of treated patients receiving X where all levels of physician decisions are confounded. This scenario illustrates some of the fundamental benefits and hazards of fee-for-service medicine: creating a financial advantage for performing a health care service results in provision of that service to more patients in need, but at the penalty of more patients receiving unnecessary services.

Note that in scenario I those patients falsely diagnosed as condition A will be viewed by their physicians as truly having condition A. Furthermore, the documentation (including the recording and interpreting of tests) will be consistent with this false diagnosis. Thus, these distortions will not be

recognizable on review of the medical records (since these are not a gold standard independent of the physician). It may not be surprising, therefore, that studies on medical practice variation which rely on medical records have failed to consistently link high utilization of medical service to "inappropriate" use of services (Chassin, 1987).

Volume and Selective Incentives

In scenario II the number of patients seen without condition A is increased to 30,000, so that the prevalence of condition A among all patients seen is now .25 instead of .5. Thus, as in the typical physician's practice, no particular diagnosis constitutes a majority of the patients seen. Note that the total number of patients seen has been increased without a change in the number of patients needing treatment. Figure II A illustrates the scenario without the financial incentive for X over Y, while figure II B illustrates the results with a financial incentive for treatment X (as in figure I B). Scenario II illustrates the fact that when diagnostic accuracy is imperfect, expanding the proportion of patients considered as potential candidates for treatment increases the total number of treatments provided, even when the total number of patients truly needing treatment is unchanged. Thus, in the face of a financial incentive for at least one of the prospective treatments, there may be an incentive to expand the population of patients seen. This demonstrates in simplest form another potential consequence of

fee-for-service medicine; financial incentives for specific services may stimulate the physician to see a larger volume of patients, in a quest for patients in need of those services.

Diagnostic Accuracy and Selective Incentives

In scenario III, the diagnostic accuracy of the physician is also reduced (from a baseline sensitivity of .9 and specificity of .9 to a new sensitivity of .8 and a specificity of .8). Selective financial incentives are assumed to alter the new baseline to a sensitivity of .85 and a specificity of .75. Thus, this scenario reduces the overall diagnostic performance of the physician (shifts the physician to an ROC curve with a smaller area) but maintains a 5 in 100 increase in sensitivity (with off-setting 5 in 100 decrease in specificity) under circumstances of financial incentive. With the lower diagnostic accuracy in IIIB, physicians perform many more of the profitable X's than they do on an identical patient population in scenario IIB. This scenario suggests that so long as the physician sees more non A patients than A patients, a financial incentive for intervention "X" may reward reduced diagnostic accuracy for condition A.

Utilization Management

These first three scenarios illustrate some of the potential consequences of selective financial incentives for physician services; increase in needed treatments, increase in unnecessary

treatments, increase in patient volume, decrease in diagnostic precision, and a shift in the mix of services provided disproportionate to the proven efficacy of those services. Indeed, in each scenario, the imposition of selective incentives for X results in a decline in the proportion of patients correctly managed (note, however, that this finding is highly dependent on the actual sensitivity and specificity). While nonselective financial incentives will correct the imbalance of services and may improve the overall appropriateness of care, the effects of incentives on increased volume and overdiagnosis (e.g. "intensity") remain for any pay-for-service scheme.

A variety of utilization management strategies have been devised for imposition on a fee-for-service reimbursement system, largely in an effort to ameliorate these effects. We will now model the effect of the most straight-forward of these utilization management interventions, that of mandatory second opinion. In this scenario (IV A and B), the payor has identified a specific service with high utilization, in this case service X. Service X is then targeted for mandatory second opinion, wherein the patients agreeing to undergo X must receive an evaluation by another professional (whose diagnostic accuracy is assumed in our scenario to be similar to that of the "X" physician). Compared to similar financial incentives without the second opinion program (IIIA or IIIB) more patients with the second opinion program (IVA and IVB) are correctly managed overall, but there is an increase in the number of mistakenly untreated patients.

Note, that unlike the previous scenarios, the imposition of selective financial incentives does not result in a decline in the overall proportion of patients correctly managed. Since in these scenarios we assume that the second opinion program is applied only to service X, there is a reduction in the proportion of X's provided over Y's. The end result of this second opinion program would be a moderation in the effects of selective financial incentives on diagnostic accuracy and compliance rate (accomplished at the expense of the second opinion program and with an increase in mistakenly untreated patients). Note that although the "second opinion" program affects the rate of diagnosis and rate of patient compliance with the original recommendation, it would only affect the "physician belief rate" (i.e. be a disincentive for physicians to learn "if A then X") when the real or perceived cost of having some patients rejected for X equaled the selective gains for performing X on the approved patients. Until this circumstance pertained, the incentive for choosing "if A then X" would remain, and so would the need for the second opinion program to control utilization. Indeed, in the face of a continued selective financial incentive for X, over time more and more physicians might choose X over Y, so that the volume of X over Y might continue to grow despite the program!

Our decision rule model for clinical problem solving can thus be used not only to illustrate the effects of selective financial incentives but also to model the influence of other types of

incentives, including utilization management interventions. Our model illustrates that utilization management techniques imposed upon a system of selective financial incentives may not eliminate distortions in physician decision making, but instead may impose conflicting incentives or other systematic distortions onto existing medical practice.

The Role of Professional Leaders

A final and obvious point illustrated by our model is that we assume authoritative opinion to be balanced regarding the correctness of rules "if A then X" and "if A then Y". Clearly an imbalance in the presentation of evidence (Davidson, 1986) or opinion, could result in a meaningful shift in the likelihood of a practicing physician's choosing a decision rule. In such a case, medical practice could be substantially distorted in the direction of the decision rule favored by most authorities, despite strong contrary financial incentives to the practitioners. Interestingly, this is consistent with the observation that the annual (technology driven) increase in cost of care for HMO's appears to rise at about the same rate as that of fee-for-service practice (Luft, 1980).

Summary

In summary, we have explored the implications of selective financial incentives on clinical decision making using the model

of clinical decision rules. This decision model and the accompanying scenarios illustrate how selective financial incentives for some medical interventions may have substantial and compounding effects. Our model suggests that modest distortions of isolated physician decisions may lead to substantial distortions of practice. According to this model, selective financial incentives may result in increased interventions for those needing such interventions, but at the cost of increased rates of incorrect application of the intervention. As a consequence the total proportion of patients correctly treated might actually decline in the face of even modest effects from selective incentives.

These scenarios suggest that selective financial incentives may encourage the physician to see more patients, even if the number of patients truly needing the profitable services cannot be increased. Furthermore, contrary to the common perception that fee-for-service medicine promotes diagnostic accuracy, selective financial incentives may actually reward the physician who is less accurate in correctly diagnosing those patients who truly need intervention.

Another suggestion from these scenarios is that utilization control measures may not correct the distortions of medical practice induced by selective financial incentives. Finally, our model suggests that distortions induced by personal incentives to the sources of decision rules (i.e. clinical investigators,

medical educators, and opinion leaders) might have profound effects on medical practice.

Future Directions

There are a number of areas of research suggested by this model of financial incentives in physician decision making. Since the learning (and remembering) of decision rules plays a key role in this model, the effect of incentives on this process is of fundamental interest. Little is currently known about the factors that guide selection of knowledge resources or the derivation of decision rules from these resources. Further, although much has been written regarding the formal evaluation of the quality of the medical literature (Sackett, 1985; Fletcher, 1986; Bennett, 1987), the factors that influence informal judgments of the credibility of clinical research or authoritative statements are presently unknown.

The role of incentives in the process of medical diagnosis also warrants further investigation. Although research in signal detection clearly demonstrates the influence of financial incentives, no such research has been performed regarding diagnostic decision-making in medical practice. There are many opportunities for financial incentives to confound diagnostic processes; research into these influences should focus not only on the traditional interpretation of diagnostic tests, but also

the medical history, physical examination, and the integrated diagnostic process.

Finally, the role of physician preferences on patient behavior warrants substantial investigation. There are many possible steps in the process of accomplishing an intervention, which often go far beyond that of a simple recommendation to the patient (Solberg, 1988). There is already some evidence that the physician's personal beliefs and preferences affect what is recommended for patients (Rich, 1986).

Even the accurate description of incentives to physicians is a subject requiring considerable clarification. For example, what physicians perceive to be profitable may be more important than what is actually profitable (Epstein, 1986). Furthermore, given that the medical specialties differ greatly in personal characteristics, training, styles of decision-making, and mode of physician-patient interaction (Eisenberg, 1979), it seems likely that the incentives perceived by members of one specialty may differ from those perceived by another. Finally, opinion leaders appear to represent a distinct group (Weinberg, 1981) whose incentives may be quite different from those of practitioners (Greer, 1987); thus, the identification of opinion leaders and the clarification of their incentives may also constitute important areas of future investigation.

This simple model of medical decision making has implications for health policy as well. It suggests that such fundamental elements of medical practice as the "correct diagnosis" and the "indicated treatment" may be malleable to the incentives of the practice environment. If the goal of health policy is to improve the quality and efficiency of care, debiasing the decision making of opinion leaders and practitioners may be an important objective. In the face of inherent uncertainty, decision makers are likely to evolve toward an optimal mix of services more quickly when decisions are not confounded by selective incentives.

The practical implications of "debiasing" are not trivial. For example, a resource based relative value scale might accurately reflect the true resource inputs for service, but the scale must also reflect the perceived profitability of services if it is to serve a "debiasing" function. Furthermore, altered fee schedules do not resolve important additional problems inherent to a fee-for-service reimbursement, such as the potential incentives for increased volume and for overdiagnosis.

Conclusions

Personal incentives are a unavoidable element of human experience, and therefore of medical practice. The incentives imposed by our current fee-for-service reimbursement may induce profound distortions in the delivery of health services. The

Figure 1VB

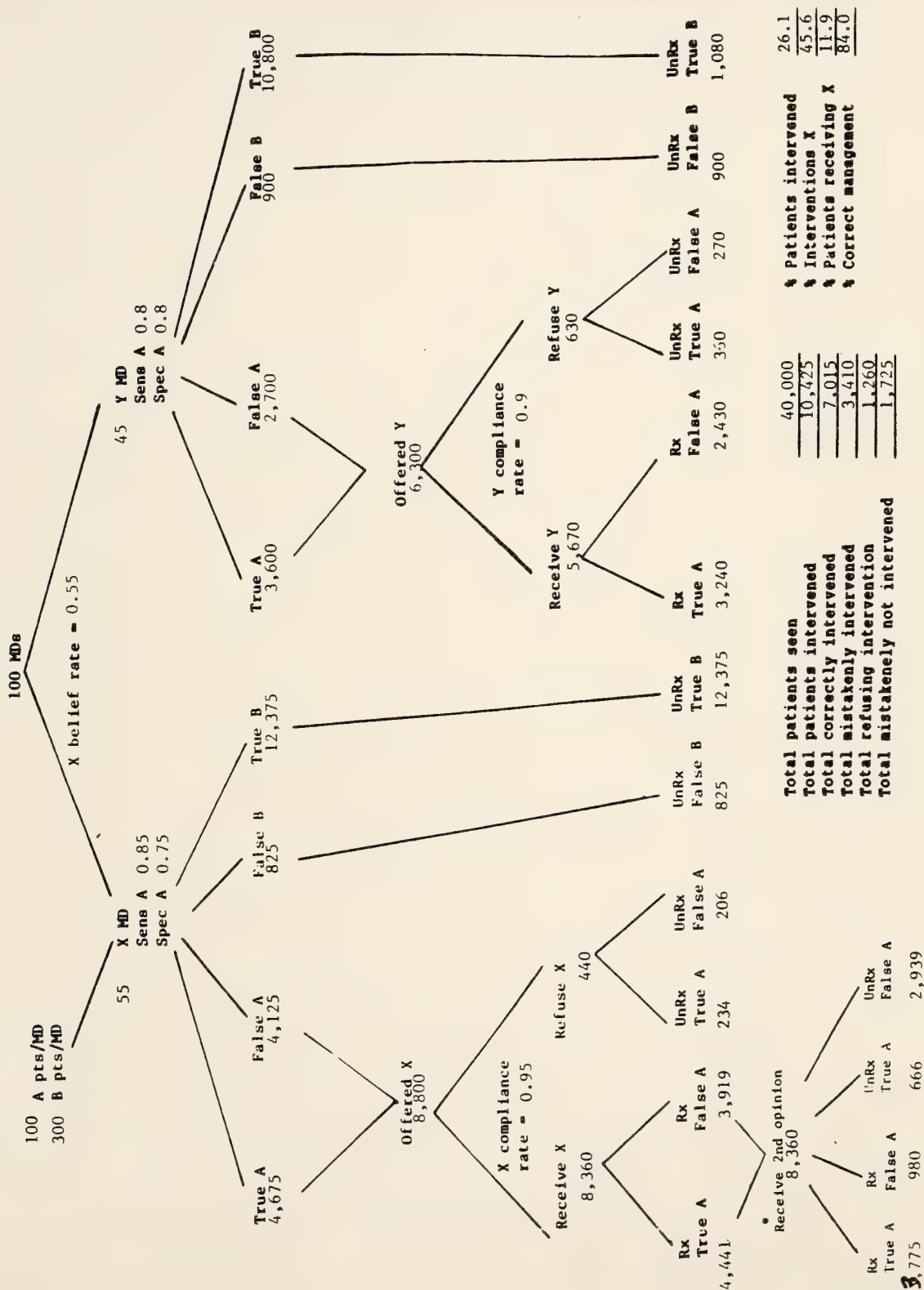


Figure IVA

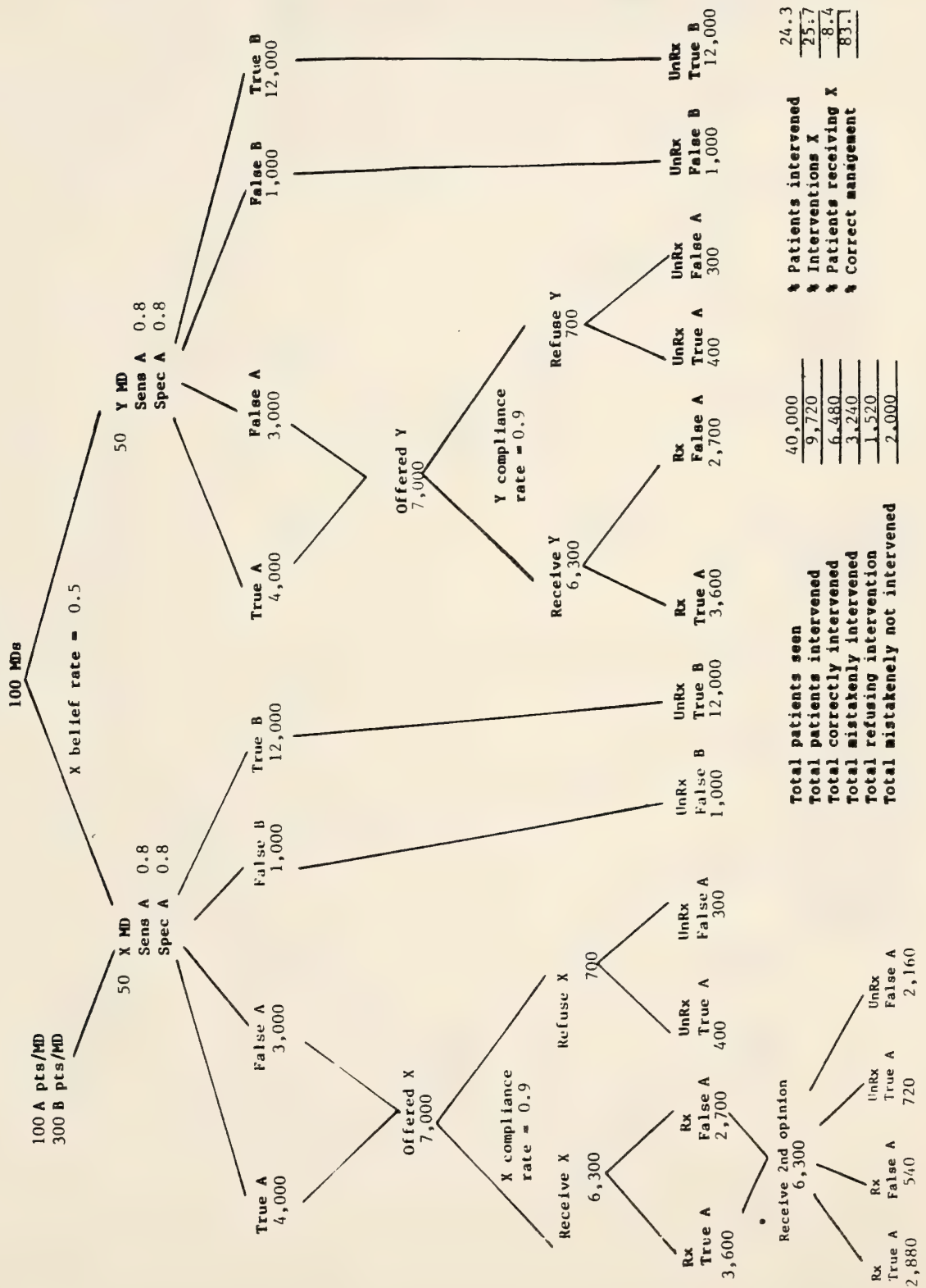


Figure IIIB

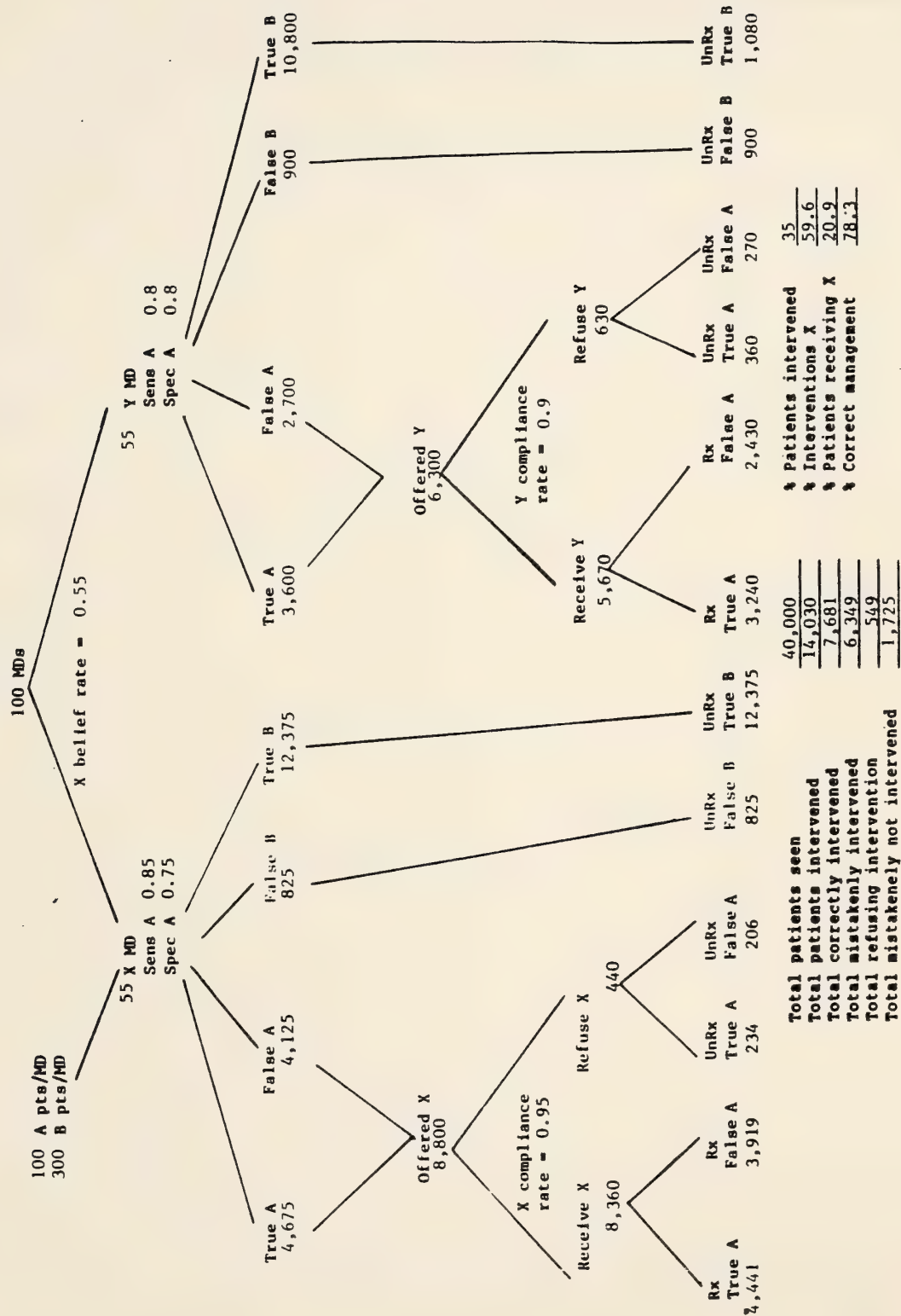


Figure IIIA

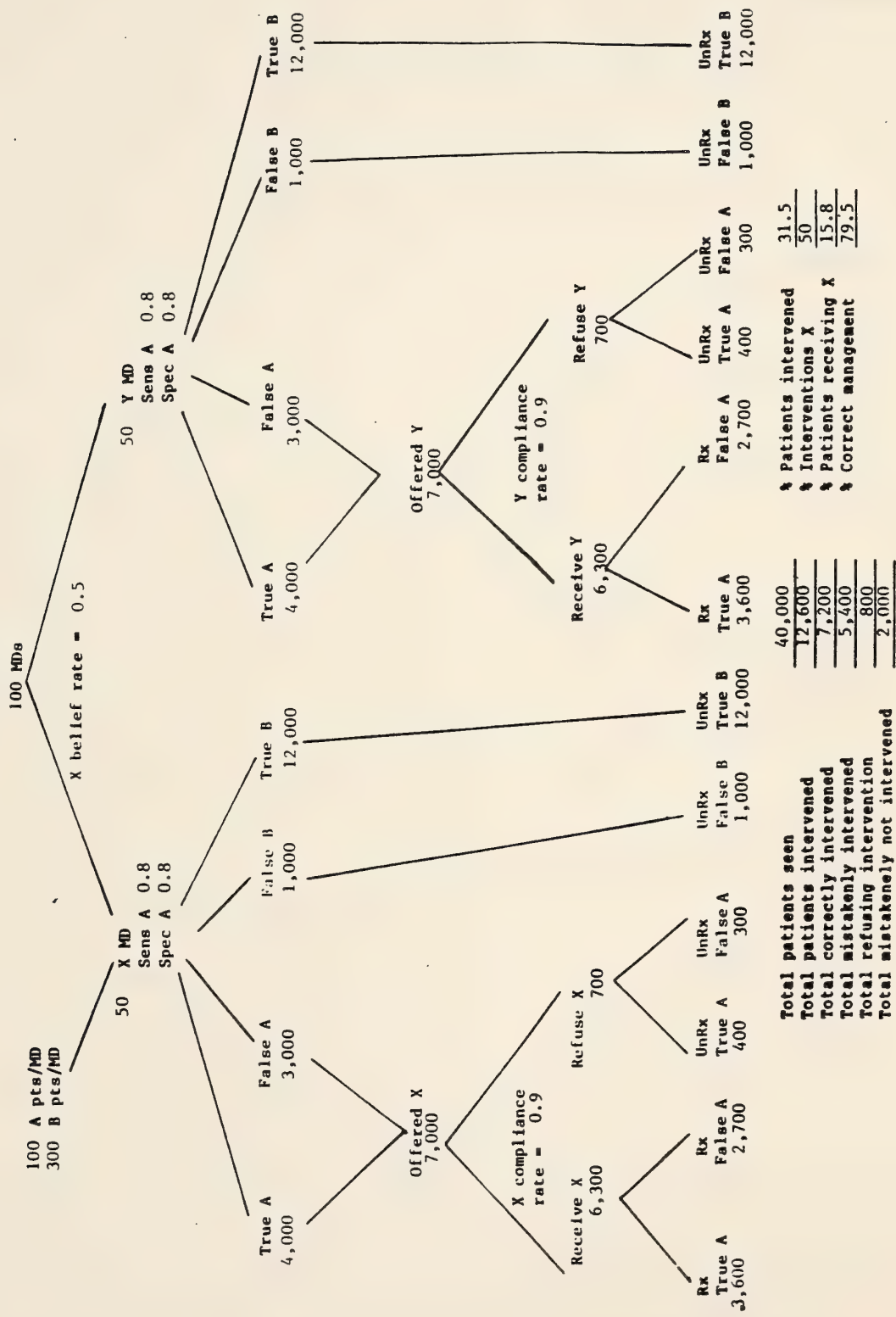


Figure IIB

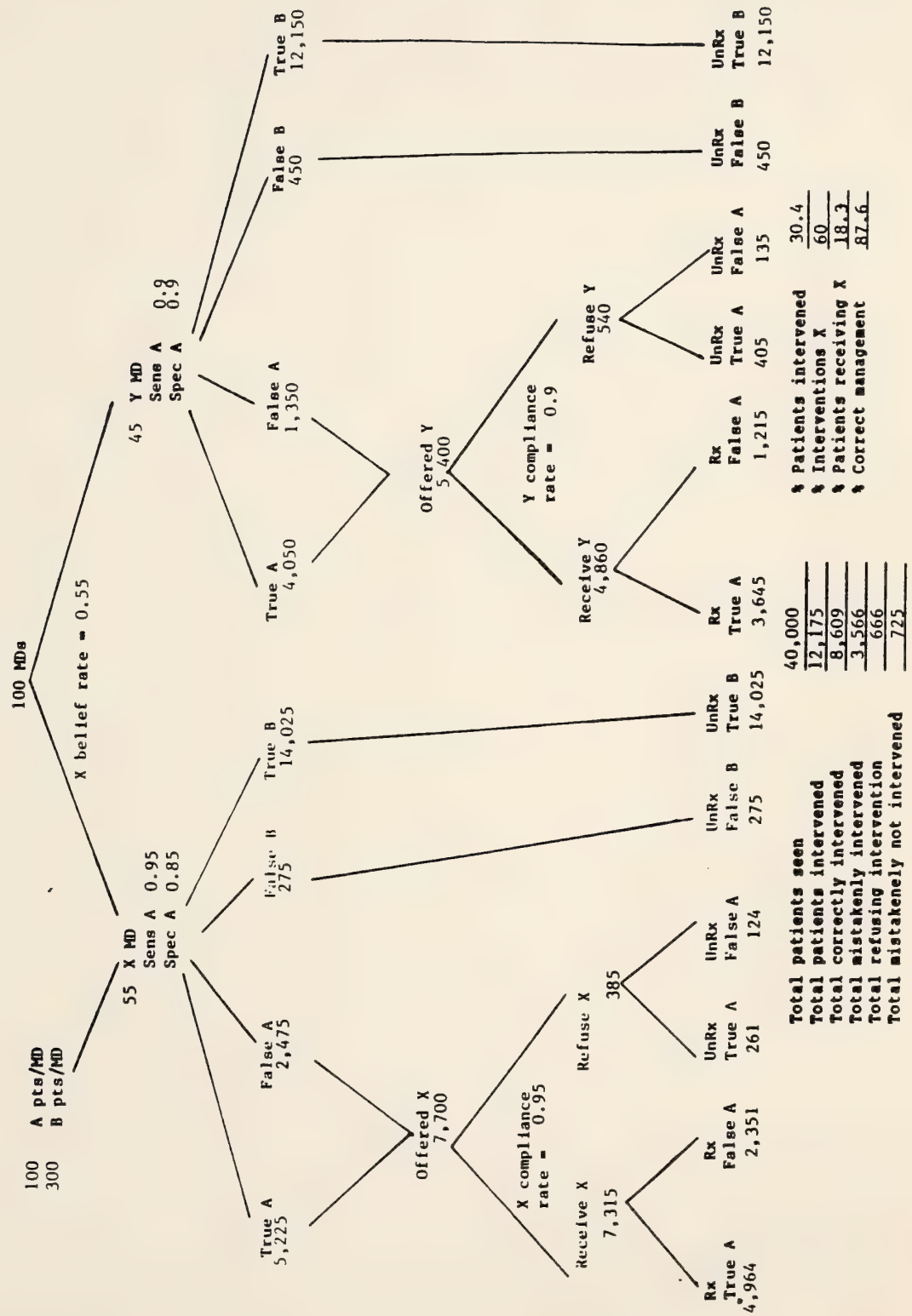


Figure IIA

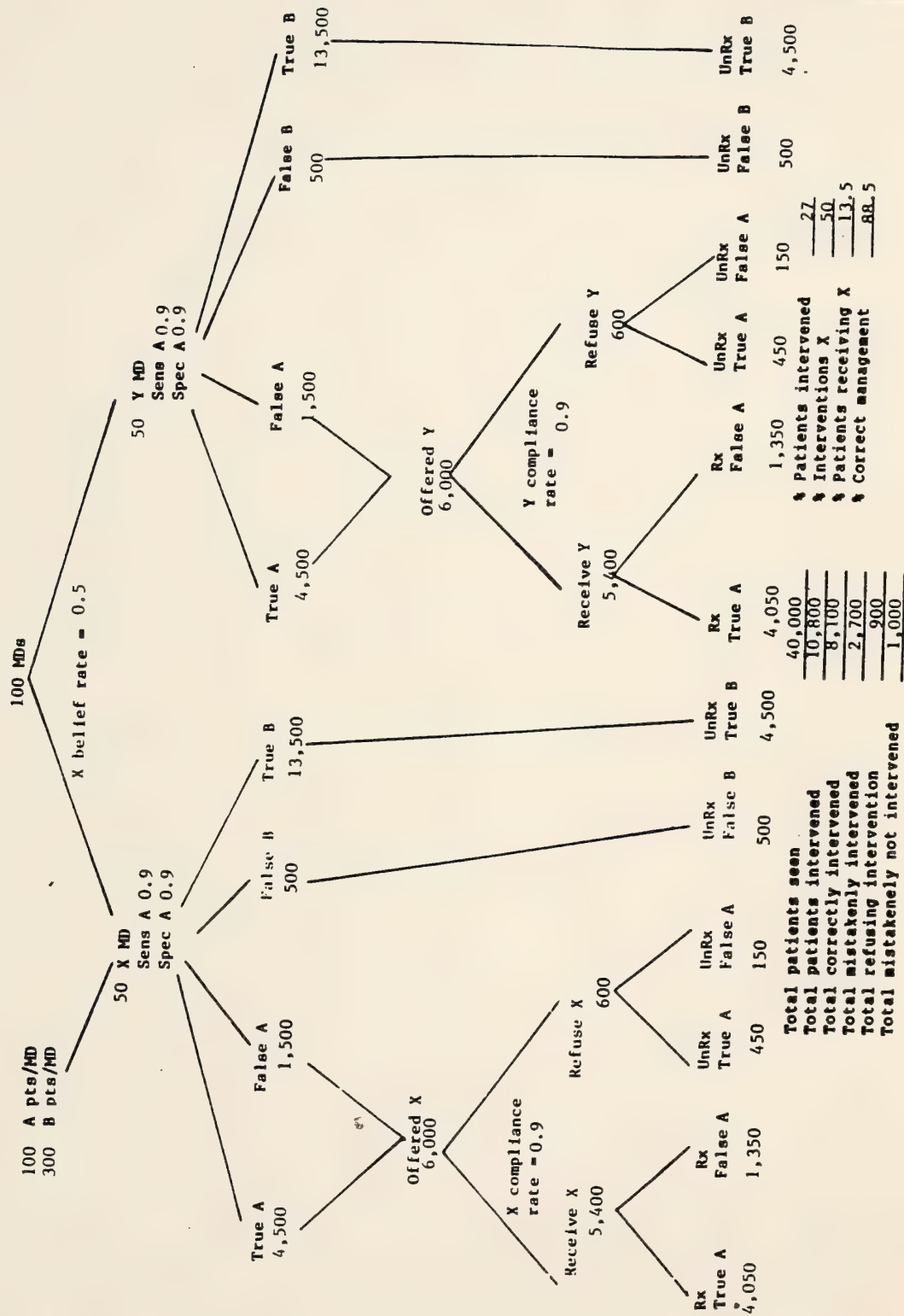


Figure 1B

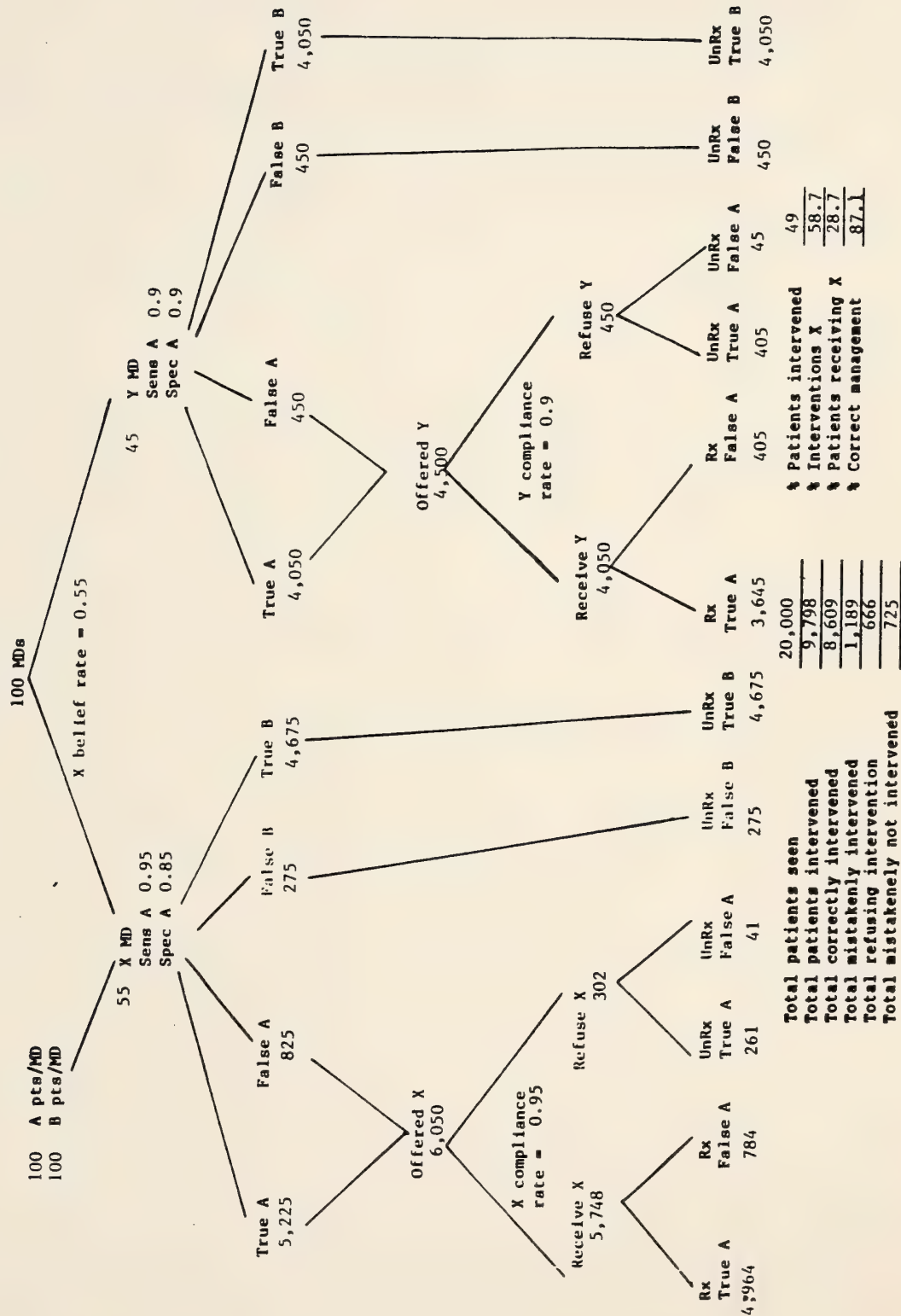
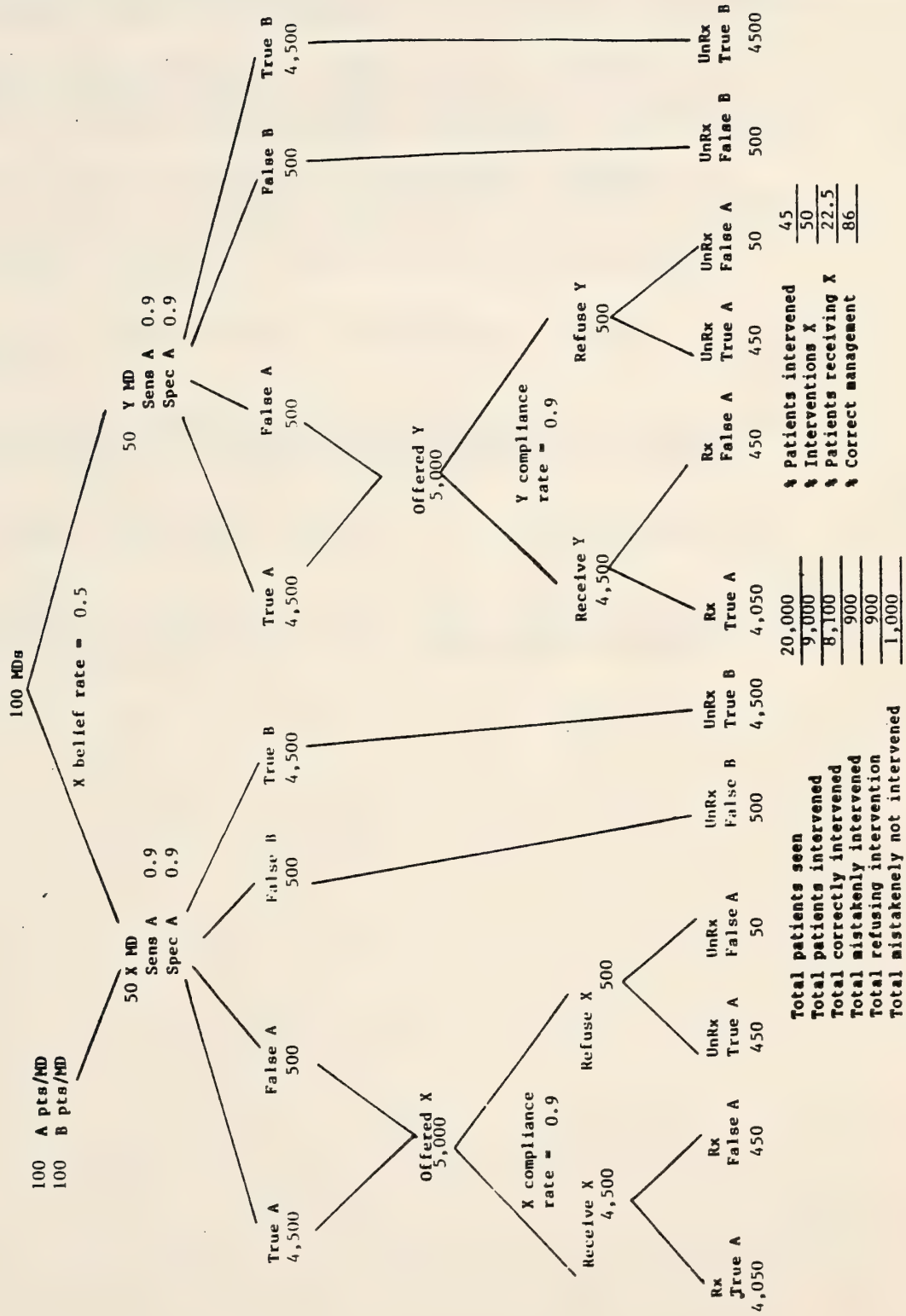


Figure 1A



Rich EC, Crowson TW, Connelly DP. Evidence for an informal clinical policy. American Journal of Medicine 1985;79:577-582.

Roper WL. Perspectives on physician payment reform. New England Journal of Medicine. 1988;319:865-867.

Sackett DL, Haynes RB, Tugwell P. How to read a clinical journal. Clinical Epidemiology A Basic Science for Clinical Medicine. Little, Brown & Company, Boston/Toronto, 1985; 285-321

Slovic P. Utility as a determiner of subjective probability. IEEE Transactions of Human Factors in Electronics 1966;7;22-28.

Solberg L, Kottke T. Doctors helping smokers. Minnesota Medicine 1988;71:413-414.

Swets JA. Measuring the accuracy of diagnostic systems. Science 1988;240:1285-1293.

Waller WS, Mitchell TR. The effect of context on the selection of diagnostic strategies. Organizational Behavior and Human Performance 1984;33:397-413.

Webster's New World Dictionary New York, NY: Warner Books Inc, 1979.

Weinberg AD, Ullian L, Richard WD, Cooper P. Informal advice and information seeking between physicians. Journal of Medical Education 1981;56:174-180.

International Journal Technology Assessment in Health Care
1988;4:5-26

Greer AL. The two cultures of biomedicine: can there be consensus? American Medical Association Journal 1987;258:2739-2740.

Hartley RM et al. Differences in ambulatory test ordering in England and America. American Journal of Medicine 1987;82:513-515.

Haynes De Regt R, Minkoff H, Feldman J, Schwartz R. Relation of private or clinic care to the cesarean birth rate. New England Journal of Medicine 1986;315:619-624.

Hickson GB, et al. Physician reimbursement by salary or fee-for-service. Pediatrics 1987;80:344-350.

Hillman AL. Financial incentive for physicians in HMO's. New England Journal of Medicine 1987;317:1743-1748.

Hlatky MA, Lee KL et al. Diagnostic test use in different practice settings. Archives of Internal Medicine 1983;143:1886-1889.

Korn JE, Schlossberg L, Rich EC. Improved preventive care after an ambulatory care rotation, carryover of a successful intervention to improve preventive practice. Journal of General Internal Medicine 1988;3:156-160

Lee W. Gambling behavior. Decision Theory and Human Behavior New York: Wiley. 1971:109-125.

Levinson DF. Toward full disclosure of referral restrictions and financial incentives by prepaid health plans. New England Journal of Medicine 1987;317:1729-1734.

Luft HS. Trends in medical care costs: do HMO's lower the rate of growth. Medical Care 1980;18:1.

McCarty DJ. Why are today's medical students choosing high-technology specialties over internal medicine? New England Journal of Medicine 1987;317:567-569.

Physician Payment Review Commission (PPRC). Medicare physician payment: an agenda for reform. Annual Report to Congress. Washington D.C: Government Printing Office. 1987:27.

Rich EC, Korn J, Schlossberg L. Effects of a preventive medicine educational program on physician attitudes and practices relevant to colon cancer screening. Clinical Research 1986;34:1014A.

Reference

- Arkes HR, Dawes RM, Christensen C. Factors in influencing the use of a decision rule. Organizational Behavior and Human Decision Processes 1986;37:93-110.
- Bennett KJ, Sackett DL, Haynes RB et al. A controlled trial of teaching critical appraisal of the clinical literature to medical students. Journal of American Medical Association 1987;257:2451-2454.
- Chassin MR, Kosecoff J, Winslow CM, et al. Does inappropriate use explain geographic variations in the use of health care services? 1987;258:2533-2537.
- Covell DG, et al. Information needs in office practice. Annals of Internal Medicine 1985;102:596-599.
- Curry L, Putnam R. Continuing medical education in maritime Canada. Canadian Medical Association Journal 1981;124:563-566.
- Davidson RA. Funding sources related to outcomes of clinical trials. Journal of General Internal Medicine 1986;1:155.
- Dawson NV, Arkes HR. Systematic errors in medical decision-making. Journal of General Internal Medicine 1987;2:1838-187.
- Eddy DM. Clinical policies and the quality of clinical practice. New England Journal of Medicine 1987;307:343-347.
- Eisenberg JM. Sociologic influences on decision-making by clinicians. Annals of Internal Medicine 1979;90:957-964.
- Epstein AM, Colin BB, McNeil BJ. The use of ambulatory testing in prepaid and fee-for-service group practices. The New England Journal of Medicine 1986;314:1089-1094.
- Festinger L, Carlsmith J. Cognitive consequences of forced compliance. Journal of Abnormal and Social Psychology 1959;58:201-210.
- Fletcher SW, Siscovick DS, Inui TS. Research on the periodic health examination. Journal of General Internal Medicine 1986;1S:S45-S49.
- Francis AM, et al. Care of patients with colorectal cancer. Medical Care 1984;22:418.
- Green D, Swets J. Signal Detection Theory and Psychophysics. New York: Krieger, 1966.
- Greer AL. The state of the art versus the state of the science: the diffusion of new medical technologies into practice.

mechanisms and effectiveness of such influences are overdue for scrutiny by researchers, policy makers, and professional leaders. The decision rule model presented here may provide a simple tool for analyzing how medical decisions might be influenced by such incentives. This model emphasizes that "professional judgment" should not be presumed to prevail over organizational interventions into the practice of medicine. Instead "professional judgment" is likely to incorporate, in potentially unanticipated ways, the incentives imposed by such interventions.

Appendix II: Survey Instruments

SURVEY INSTRUMENT:
MEDICARE CARRIERS

SEP 12 1983

NOTE TO: All Part B Carriers

SUBJECT: Congressionally Mandated Study on the Effectiveness of Medical Review—
ACTION

Section 4056(c)(2) of the Omnibus Budget Reconciliation Act (OBRA) of 1987 requires the Secretary to study a number of issues related to volume and intensity of physicians' services. One issue of particular relevance to both you and us is an analysis of the effectiveness of methods currently used in Medicare to ensure that payments are made only for medically necessary services.

We have decided to utilize the services of a research center that currently has a cooperative agreement with HCFA to assist us in the required study. The center, a part of the school of Public Health at the University of Minnesota is also conducting the remaining studies relating to volume and intensity required under 4056(c)(2), for the Office of Research and Demonstrations.

The attached survey has been prepared by research center staff. I urge your cooperation in completing the survey.

The current project plan calls for several site visits by Minnesota staff. We will be discussing potential sites with the regional offices. If a visit is proposed to your organization, we will provide specific information on what will need to be accomplished as far in advance of the visit as possible.

I appreciate your cooperation in working with us to ensure that the University of Minnesota staff can successfully complete their study in a timely manner.



Carol J. Walton
Director
Office of Program Operations Procedures

Attachment

cc:
Associate Regional Administrators
for Program Operations
Regions I, II, IV-VII, IX-X
Associate Regional Administrators
for Medicare
Regions II, VII



UNIVERSITY OF MINNESOTA
TWIN CITIES

Division of Health Services Research and Policy
420 Delaware Street S.E., Box 729
Minneapolis, Minnesota 55455-0392
(612) 624-6151

SURVEY OF CARRIER VOLUME AND INTENSITY CONTROLS
FOR PART B OF MEDICARE

Please Read This Page First

Public Law 100-203 requires the Secretary of the Department of Health and Human Services to analyze the effectiveness of methods currently used to ensure that Medicare payments are made only for medically necessary physicians' services. To help fulfill this mandate, the Health Care Financing Administration (HCFA) has awarded a cooperative agreement to the University of Minnesota to study methods used by carriers to control the volume and intensity of physicians' services. The University of Minnesota, in cooperation with HCFA staff, has designed this survey. You, and each of the other carriers in the country, have received a copy of the survey.

We would greatly appreciate your help in completing this survey. The questions are detailed and specific; it would be most helpful if you assigned the task of completing the survey to staff that are knowledgeable in Medicare utilization review.

Instead of returning the survey to us right away, we would like you to fill it out and hold it. Colleen Grogan, from the University of Minnesota, will call you in about one week to go over the survey. She will be available to review any problem areas with you and to write down your answers. We believe that this is the most accurate method to obtain the necessary information.

After Colleen has completed her interview, could you please return the survey (or a copy of it, if you wish to keep the original) to:

Roger Feldman, Ph.D.
HCFA Research Center Director
Division of Health Services Research and Policy
University of Minnesota
420 Delaware Street, S.E., Box 729
Minneapolis, MN 55455

Our phone number is 612-624-6151. Please call us if you have any questions. Can we also have the name of a contact person in your organization?

Name: _____ Phone: _____

SURVEY OF CARRIER VOLUME AND INTENSITY CONTROLS
FOR PART B OF MEDICARE

The purpose of this survey is to collect information on the Medical Review (MR) activities of Medicare Part B Carriers. Medical Review consists of pre- and postpayment review to identify inappropriate, medically unnecessary, or excessive services.

PART I: PREPAYMENT SCREENS

A. HCFA Mandated Prepayment Screens and Parameters

1. We would like to know when you implemented each of the 13 mandated screens, and whether you use tighter screening parameters than those mandated by HCFA.

<u>SCREEN</u>	<u>DATE IMPLEMENTED</u> (mo/year)	<u>MANDATED PARAMETER</u>	<u>YOUR PARAMETER</u> (if tighter)
Routine foot care	_____	1 treatment per 60 days	_____
Mycotic nails	_____	1 treatment per 60 days	_____
Nursing home visits	_____	1 visit per month	_____
New patient office visits	_____	1 comprehensive physical exam per carrier history period	_____
Holter and real-time monitoring	_____	1 instance per 6 months	_____
Chiropractic	_____	12 spinal manipulations per year	_____
Concurrent care	_____	1 doctor of same specialty billing for in-hospital services on same day	_____
Hospital visits	_____	31 times in 3 months	_____
Comprehensive office visits	_____	1 per six months	_____

<u>SCREEN</u>	<u>DATE IMPLEMENTED</u> (mo/year)	<u>MANDATED PARAMETER</u>	<u>YOUR PARAMETER</u>
SNF visits	_____	2 subsequent care visits in 1st week, 1 visit per week thereafter	_____ _____
Injections	_____	24 per year	_____
Urological supplies	_____	2 catheters per month	_____
Replacement of post-cataract external prosthetic contact lenses	_____	1 per eye per year	_____

2. What is the skill level of the person who normally performs medical review for mandated screens? (check one)

Claims examiner with limited medical training _____

Nurse _____

Physician _____

Other _____

If other, please identify: _____

2a. What factors determine the skill level needed for a particular type of review?

2b. What was the impact of your budget in determining the skill level needed to conduct medical review?

3. What is the estimated cost of mandated prepayment review activities in each of the three most recent fiscal years?

FY from _____ to _____ \$ _____

FY from _____ to _____ \$ _____

FY from _____ to _____ \$ _____

- 3a. Do you keep more-detailed records of the cost of mandated prepayment review, e.g., by 3-month period or by type of service reviewed? (check one)

Yes _____

No _____

If yes, please explain: _____

4. Can you describe any problems encountered in implementing the mandated prepayment screens (e.g., with providers)?

5. We are interested in your assessment of the non-quantifiable cost-avoidance effects of mandated prepayment screens. Could you answer the following questions, even if you have to estimate the answers: What percentage of the unscreened claims for each of the following screens are medically unnecessary (MU)? Did simple presence of the screen cause the percentage of unscreened claims that are medically unnecessary to increase or decrease?

<u>SCREEN</u>	<u>% UNSCREENED CLAIMS "MU"</u>	<u>EFFECT OF SCREEN (CHECK ONE):</u>			
		INCREASE	DECREASE	NO CHANGE	DON'T KNOW
Routine foot care	_____ %	_____	_____	_____	_____
Mycotic nails	_____ %	_____	_____	_____	_____
Nursing home visits	_____ %	_____	_____	_____	_____

<u>SCREEN</u>	<u>% UNSCREENED CLAIMS "MU"</u>	<u>EFFECT OF SCREEN (CHECK ONE):</u>			
		INCREASE	DECREASE	NO CHANGE	DON'T KNOW
New pat. office visits	_____ %	_____	_____	_____	_____
Holter monitor	_____ %	_____	_____	_____	_____
Chiropractic	_____ %	_____	_____	_____	_____
Concurrent care	_____ %	_____	_____	_____	_____
Hospital visits	_____ %	_____	_____	_____	_____
Comprehensive office vis.	_____ %	_____	_____	_____	_____
SNF visits	_____ %	_____	_____	_____	_____
Injections	_____ %	_____	_____	_____	_____
Urological supplies	_____ %	_____	_____	_____	_____
Contact lens replacement	_____ %	_____	_____	_____	_____

B. Other Prepayment Screens

1. A medical or surgical procedure must be suspended for manual review when the submitted charge exceeds a predetermined percent of the prevailing charge (250% is suggested). What percent do you use for this screen? %
2. "Once-in-a-lifetime" procedures must be identified. Please indicate the HCPCS codes of such procedures that you have identified: (if the list is too long, you may attach a copy)

3. Many carriers have developed optional screens, in addition to those mandated by HCFA. We are interested in the top 10 optional screens that you use, as measured by dollar volume of claims screened. On the following page we ask a series of questions about the most important optional screen. Please make 9 copies of this page if you have 10 optional screens, or make the appropriate number of copies if you have fewer than 10 optional screens. After completing these pages, please continue to the next set of questions.

OPTIONAL PREPAYMENT SCREENS

1. Description of screen: _____
 HCPCS code(s): _____
 Date implemented: _____
 Screening parameter: _____

2. What is the skill level of the person who normally performs the medical review?
 Claims examiner with limited medical training _____
 Nurse _____
 Physician _____
 Other _____
 If other, please identify: _____

3. Does the screen result in (check one):
 Automatic denial? _____
 Manual review? _____

4. What is the estimated cost of this screen during the 3 most recent fiscal years?
 FY from _____ to _____ \$ _____
 FY from _____ to _____ \$ _____
 FY from _____ to _____ \$ _____

5. Please describe any problems encountered in implementing this screen (e.g., with providers): _____

6. What percent of unscreened claims are medically unnecessary? ____ %
 What was the effect of this screen on the percent of "MU"
 unscreened claims? INCREASE ____ DECREASE ____ NO CHANGE ____ DK ____

7. For the 3 most recent fiscal years, please indicate:

	MOST RECENT YEAR	2nd YEAR	3rd YEAR
Number of claims screened	_____	_____	_____
\$ volume of claims screened	\$ _____	\$ _____	\$ _____
No. services denied/reduced	_____	_____	_____
\$ services denied/reduced	\$ _____	\$ _____	\$ _____
No. denials/reductions reversed	_____	_____	_____
\$ denials/reductions reversed	_____	_____	_____

..... PLEASE MAKE EXTRA COPIES AS NEEDED

C. Relative Effectiveness of Mandated Versus Optional Prepayment Screens

1. We are interested in your assessment of the relative effectiveness of mandated versus optional pre-payment screens. Are there any mandated screens that, in your opinion, are not cost-effective? If so, why not? Please attach additional pages if the following blanks are not adequate:

1a. First ineffective mandated screen: _____

Reason: _____

1b. Second ineffective mandated screen: _____

Reason: _____

1c. Third ineffective mandated screen: _____

Reason: _____

2. Would you recommend adding any of your optional screens to the list of mandated national screens, based on their superior cost-effectiveness? If so, which ones?

2a. First recommended screen: _____

2b. Second recommended screen: _____

2c. Third recommended screen: _____

3. Do you think that local variation in medical review policies and guidelines impacts the relative effectiveness of both national and optional screens? Please discuss.

PART II: POSTPAYMENT SCREENS

A. General Questions

1. What is the skill level of the person who normally performs postpayment MR? (check one)

Claims examiner with limited medical training _____
Nurse _____
Physician _____
Other _____
If other, please identify: _____

- 1a. What factors determine the skill level needed for postpayment MR?

2. Please describe the peer groups used for physician profiling in your postpayment MR system, e.g., peer groups may be defined by specialty, subspecialty, locality, or other methods:

3. After you have identified physician/suppliers who need further investigation, what types of educational activities are done? Please describe these activities and the order you do them:

B. Specific Postpayment Pattern-of-Practice Comparisons

Carriers are required to make postpayment pattern-of-practice comparisons by calculating two ratios: Ratio I is the number of services provided by a physician/supplier per 100 beneficiaries. Ratio II is the number of services per each beneficiary who actually received services in that category. Examples are:

Ratio I
Total # of EKGs = 15
Total # of patients seen = 84
15 divided by 84 times 100 = 17.9
Therefore, 18 EKGs were
performed per 100 patients

Ratio II
Total # of EKGs = 15
Total # of patients receiving EKGs = 7
15 divided by 7 = 2.14
Therefore, 2 EKGs were performed on
each patient who received an EKG

What was the critical value for each ratio that would cause a physician/supplier to be selected for postpayment review?

<u>Category</u>	<u>Ratio I</u>	<u>Ratio II</u>
Office visits	_____	_____
Home visits	_____	_____
Hospital visits	_____	_____
SNF visits	_____	_____
Nursing home visits	_____	_____
Injections	_____	_____
EKGs	_____	_____
Surgery	_____	_____
Office lab services	_____	_____
Office diagnostic x-ray	_____	_____
Physical therapy	_____	_____
Other comparisons (e.g., oxygen concentrators, seat lift chairs, or air ambulance:		
_____	_____	_____
_____	_____	_____
_____	_____	_____

PART III: OTHER METHODS OF CONTROLLING VOLUME AND INTENSITY

- A. In an attempt to control the number and cost of physician services, carriers have implemented changes in their payment structure. Have you tried changing the coding of physician services to prevent physicians from "upcoding" procedures (e.g., have you reduced the number of office visit codes you recognize)?

yes (continue)

no (skip to B)

1. Could you describe how the coding of physician services was changed?

2. When were the coding changes implemented? _____

3. Have you formally evaluated the effectiveness of the coding changes?

yes (continue)

no (skip to A.9)

don't know (skip to A.9)

4. Could you describe this evaluation program? _____

5. What has been the effect of the coding changes on total cost?

increase by _____ %

decrease by _____ %

no change _____

don't know _____

6. What has been the effect of the coding changes on cost of physician services?

increase by _____ %
decrease by _____ %
no change _____
don't know _____

7. What has been the effect of the coding changes on total utilization?

increase by _____ %
decrease by _____ %
no change _____
don't know _____

8. What has been the effect of the coding changes on utilization of physician services?

increase by _____ % (skip to B)
decrease by _____ % (skip to B)
no change _____ (skip to B)
don't know _____ (skip to B)

9. What is your overall feeling toward the effectiveness of the coding changes?

10. How do you think the coding changes have affected the cost and utilization of physician services?

B. Do you extensively "bundle" physician services into broader categories (e.g., global fees for surgery)?

yes _____ (continue)
no _____ (skip to C)

1. Could you describe the bundling program?

2. When was the bundling program implemented? _____

3. Have you formally evaluated the effectiveness of the bundling program?

yes _____ (continue)
no _____ (skip to B.9)
don't know _____ (skip to B.9)

4. Could you describe this evaluation program? _____

5. What has been the effect of the bundling program on total cost?

increase by _____ %
decrease by _____ %
no change _____
don't know _____

6. What has been the effect of the bundling program on cost of physician services?

increase by _____ %
decrease by _____ %
no change _____
don't know _____

7. What has been the effect of the bundling program on total utilization?

increase by _____ %
decrease by _____ %
no change _____
don't know _____

8. What has been the effect of the bundling program on utilization of physician services?

increase by _____ % (skip to C)
decrease by _____ % (skip to C)
no change _____ (skip to C)
don't know _____ (skip to C)

9. What is your overall feeling toward the effectiveness of the bundling program?

10. How do you think the bundling program has affected the cost and utilization of physician services?

C. Describe other techniques for controlling volume and intensity that you have successfully used under part B of Medicare:

PART IV: SPECIAL SECTION FOR AMBULATORY CARDIAC MONITORING

Prior to September 1987, the CPT-4 codes for certain types of cardiac monitoring did not provide carriers with an adequate basis for accumulating charge data and recognizing price differentials in establishing reasonable charge allowances. In response to these problems, the Director of the Bureau of Eligibility, Reimbursement and Coverage established several temporary alpha-numeric codes for Holter and Real-time monitoring. For each of the cardiac monitoring codes shown below, please tell us your current prevailing charge for participating internists. If you administer payment of Part B claims in more than one geographic area, please provide the prevailing charge for each carrier region.

Codes for Holter Monitoring

Prevailing Charge

- | | |
|--|-------|
| Q 0019 Electrocardiographic monitoring for 24 hours
by continuous original ECG waveform recording
and storage with visual superimposition scanning | _____ |
| Q 0020 recording only | _____ |
| Q 0021 scanning analysis with report | _____ |
| Q 0022 physician review and interpretation | _____ |
| Q 0023 Monitoring without superimposition scanning
utilizing a printout device | _____ |
| Q 0024 recording only | _____ |
| Q 0025 microprocessor-based analysis with report | _____ |
| Q 0026 physician review and interpretation | _____ |

Codes for Real-Time Monitoring

- | | |
|---|-------|
| Q 0027 Real-time monitoring utilizing a device capable
of producing 75 or more waveform tracings | _____ |
| Q 0028 monitoring and real-time data analysis with report | _____ |
| Q 0029 physician review and interpretation | _____ |
| Q 0030 Real-time monitoring utilizing a device capable
of producing up to 75 waveform tracings | _____ |
| Q 0031 monitoring and real-time data analysis with report | _____ |
| Q 0032 physician review and interpretation | _____ |

THIS COMPLETES THE SURVEY. THANK YOU FOR YOUR COOPERATION.

SURVEY INSTRUMENT:
COMMERCIAL CARRIERS

Interviewer _____
Date _____

SECTION A

CONVENTIONAL (INDEMNITY) GROUP BUSINESS

Company Name _____
Company contact (title) _____
Phone number _____
Stratum Number _____
Company ID Number _____

The following questions are going to be about your conventional (indemnity) group business. Are you the person most knowledgeable about this topic?

(IF YES, GO TO Q1)

(IF NO, ASK FOR APPROPRIATE PERSON AND PHONE NUMBER)

1. How do you typically (most common) reimburse physicians?
(READ CATEGORIES AND CIRCLE CORRECT RESPONSE)

Billed charges..... 1
Usual and customary charges (UCR)..... 2
Discounted charges as payment in full.... 3
Fee schedule..... 4
Capitation payment..... 5
Other..... 6
(specify) _____

DK.....9

2. How do you typically (most common) reimburse hospitals?
(READ CATEGORIES AND CIRCLE CORRECT RESPONSE)

Billed charges..... 1
Usual and customary charges (UCR)..... 2
Discounted charges as payment in full.... 3
Per diem indemnity payment..... 4
DRGs or other per case payment methods... 5
Other..... 6
(specify) _____

DK.....9

Could you please tell me whether you provide the following utilization review activities?

3. pre-admission certification

YES.....1 (Go to Q3A)
NO2 (Go to Q4)
DK.....9 (Go to Q4)

- A) Based on your premium dollars, what percent of business is covered by pre-admission certification?
(INTERVIEWER: Ask for subscribers or contracts if unable to give dollars)

_____‡ (not asked = -1)

- B) In the past two years have you formally evaluated the effectiveness of pre-admission certification?

NA.....0
YES.....1 (Go to Q3C)
NO.....2 (Go to Q3E)
DK.....9 (Go to Q4)

- C) What has been the effect on total cost?

NA.....0
INC.....1 (Go to Q3Ci)
DEC.....2 (Go to Q3Cii)
NOCHANGE.....3 (Go to Q3D)
DK.....9 (Go to Q3D)

i) What is the percent increase?_____‡

ii) What is the percent decrease?_____‡

- D) What has been the effect on utilization?

NA.....0
INC.....1 (Go to Q3Di)
DEC2 (Go to Q3Dii)
NOCHANGE.....3 (Go to Q4)
DK.....9 (GO to Q4)

i) What is the percent increase?_____‡ (Go to Q4)

ii) What is the percent decrease?_____‡ (Go to Q4)

- E) i) What is your overall feeling toward the effectiveness of pre-admission certification?

ii) How do you think pre-admission certification has effected cost and utilization outside the hospital?

(And do you provide) concurrent review

YES.....1 (Go to Q4A)

NO2 (Go to Q5)

DK.....9 (Go to Q5)

A) Based on your premium dollars, what percent of business is covered by concurrent review? (INTERVIEWER: Ask for subscribers or contracts if unable to give dollars)

_____ % (not asked = -1)

B) In the past two years have you formally evaluated the effectiveness of concurrent review?

NA.....0

YES.....1 (Go to Q4C)

NO.....2 (Go to Q4E)

DK.....9 (Go to Q5)

C) What has been the effect on total cost?

NA.....0

INC.....1 (Go to Q4Ci)

DEC.....2 (Go to Q4Cii)

NOCHANGE.....3 (Go to Q4D)

DK.....9 (Go to Q4D)

i) What is the percent increase? _____ %

ii) What is the percent decrease? _____ %

D) What has been the effect on utilization?

NA.....0

INC.....1 (Go to Q4Di)

DEC2 (Go to Q4Dii)

NOCHANGE.....3 (Go to Q5)

DK.....9 (Go to Q5)

i) What is the percent increase? _____ % (Go to Q5)

ii) What is the percent decrease? _____ % (Go to Q5)

E) i) What is your overall feeling toward the effectiveness of concurrent review?

ii) How do you think concurrent review has effected cost and utilization outside the hospital?

6. (And do you provide) retrospective review of outpatient care
YES.....1 (Go to Q6A)
NO2 (Go to Q7)
DK.....9 (Go to Q7)

A) Based on your premium dollars, what percent of business is covered by retrospective review of outpatient care (INTERVIEWER: Ask for subscribers or contracts if unable to give dollars)
_____ % (not asked = -1)

B) In the past two years have you formally evaluated the effectiveness of retrospective review of outpatient care?
NA.....0
YES.....1 (Go to Q6C)
NO.....2 (Go to Q6E)
DK.....9 (Go to Q7)

C) What has been the effect on total cost?
NA.....0
INC.....1 (Go to Q6Ci)
DEC.....2 (Go to Q6Cii)
NOCHANGE.....3 (Go to Q6D)
DK.....9 (Go to Q6D)

i) What is the percent increase?_____ %

ii) What is the percent decrease?_____ %

D) What has been the effect on utilization?
NA.....0
INC.....1 (Go to Q6Di)
DEC2 (Go to Q6Dii)
NOCHANGE.....3 (Go to Q7)
DK.....9 (Go to Q7)

i) What is the percent increase?_____ % (Go to Q7)

ii) What is the percent decrease?_____ % (Go to Q7)

E) i) What is your overall feeling toward the effectiveness of retrospective review of outpatient care?

ii) How do you think retrospective review of outpatient care has effected cost and utilization outside the hospital?

7. (And do you provide) mandatory second opinion for some surgical procedures

YES.....1 (Go to Q7A)
NO2 (Go to Q8)
DK.....9 (Go to Q8)

A) Based on your premium dollars, what percent of business is covered by mandatory second opinion for some surgical procedures? (INTERVIEWER: Ask for subscribers or contracts if unable to give dollars)
_____ % (not asked = -1)

B) In the past two years have you formally evaluated the effectiveness of mandatory second opinion for some surgical procedures?

NA.....0
YES.....1 (Go to Q7C)
NO.....2 (Go to Q7E)
DK.....9 (Go to Q8)

C) What has been the effect on total cost?

NA.....0
INC.....1 (Go to Q7Ci)
DEC.....2 (Go to Q7Cii)
NOCHANGE.....3 (Go to Q7D)
DK.....9 (Go to Q7D)

i) What is the percent increase? _____ %

ii) What is the percent decrease? _____ %

D) What has been the effect on utilization?

NA.....0
INC.....1 (Go to Q7Di)
DEC2 (Go to Q7Dii)
NOCHANGE.....3 (Go to Q8)
DK.....9 (Go to Q8)

i) What is the percent increase? _____ % (Go to Q8)

ii) What is the percent decrease? _____ % (Go to Q8)

E) i) What is your overall feeling toward the effectiveness of mandatory second opinion for some surgical procedures?

ii) How do you think mandatory second opinion for some surgical procedures has effected cost and utilization outside the hospital?

8. (And do you provide) high cost case management

YES.....1 (Go to Q8A)
NO2 (Go to Q9)
DK.....9 (Go to Q9)

A) Based on your premium dollars, what percent of
business is covered by high cost case management
(INTERVIEWER: Ask for subscribers or contracts if unable
to give dollars)

_____ % (not asked = -1)

B) In the past two years have you formally evaluated the
effectiveness of high cost case management?

NA.....0
YES.....1 (Go to Q8C)
NO.....2 (Go to Q8E)
DK.....9 (Go to Q9)

C) What has been the effect on total cost?

NA.....0
INC.....1 (Go to Q8Ci)
DEC.....2 (Go to Q8Cii)
NOCHANGE.....3 (Go to Q8D)
DK.....9 (Go to Q8D)

i) What is the percent increase?_____ %

ii) What is the percent decrease?_____ %

D) What has been the effect on utilization?

NA.....0
INC.....1 (Go to Q8Di)
DEC2 (Go to Q8Dii)
NOCHANGE.....3 (Go to Q9)
DK.....9 (Go to Q9)

i) What is the percent increase?_____ % (Go to Q9)

ii) What is the percent decrease?_____ % (Go to Q9)

E) i) What is your overall feeling toward the effectiveness
of high cost case management?

ii) How do you think high cost case management has
effected cost and utilization outside the hospital?

9. (And do you provide) physician profiling and feedback

YES.....1 (Go to Q9A)

NO2 (Go to Q10)

DK.....9 (Go to Q10)

A) Based on your premium dollars, what percent of business covered by physician profiling and feedback (INTERVIEWER: Ask for subscribers or contracts if unable to give dollars)

_____ % (not asked = -1)

B) In the past two years have you formally evaluated the effectiveness of physician profiling and feedback?

NA.....0

YES.....1 (Go to Q9C)

NO.....2 (Go to Q9E)

DK.....9 (Go to Q10)

C) What has been the effect on total cost?

NA.....0

INC.....1 (Go to Q9Ci)

DEC.....2 (Go to Q9Cii)

NOCHANGE.....3 (Go to Q9D)

DK.....9 (Go to Q9D)

i) What is the percent increase? _____ %

ii) What is the percent decrease? _____ %

D) What has been the effect on utilization?

NA.....0

INC.....1 (Go to Q9Di)

DEC2 (Go to Q9Dii)

NOCHANGE.....3 (Go to Q10)

DK.....9 (Go to Q10)

i) What is the percent increase? _____ % (Go to Q10)

ii) What is the percent decrease? _____ % (Go to Q10)

E) i) What is your overall feeling toward the effectiveness of physician profiling and feedback?

ii) How do you think physician profiling and feedback effected cost and utilization outside the hospital?

10. (And do you provide) discharge planning

YES.....1 (Go to Q10A)

NO2 (Go to Q11)

DK.....9 (Go to Q11)

A) Based on your premium dollars, what percent of business is covered by discharge planning?

(INTERVIEWER: Ask for subscribers or contracts if unable to give dollars)

_____ % (not asked = -1)

B) In the past two years have you formally evaluated the effectiveness of discharge planning?

NA.....0

YES.....1 (Go to Q10C)

NO.....2 (Go to Q10E)

DK.....9 (Go to Q11)

C) What has been the effect on total cost?

NA.....0

INC.....1 (Go to Q10Ci)

DEC.....2 (Go to Q10Cii)

NOCHANGE.....3 (Go to Q10D)

DK.....9 (Go to Q10D)

i) What is the percent increase? _____ %

ii) What is the percent decrease? _____ %

D) What has been the effect on utilization?

NA.....0

INC.....1 (Go to Q10Di)

DEC2 (Go to Q10Dii)

NOCHANGE.....3 (Go to Q11)

DK.....9 (Go to Q11)

i) What is the percent increase? _____ % (Go to Q11)

ii) What is the percent decrease? _____ % (Go to Q11)

E) i) What is your overall feeling toward the effectiveness of discharge planning?

ii) How do you think discharge planning has effected cost and utilization outside the hospital?

11. (IF THEY USE ONE OR MORE OF THE UTILIZATION REVIEW TECHNIQUES ABOVE, ASK:) How do you evaluate their effectiveness?

12. (IF THEY DO NOT USE ONE OR MORE OF THE UTILIZATION REVIEW TECHNIQUES ABOVE, ASK) Please indicate any reason(s) you may have for not using some of the utilization review techniques mentioned before. (PROBE BY REPEATING TECHNIQUES THEY DO NOT USE)

13. For the year 1988, what percent of your group health business, in premium dollars, including self-insured business, is: (INTERVIEWER: Ask for subscribers or contracts if unable to give dollars)(INTERVIEWER: DON'T ASK EITHER FEE-FOR-SERVICE PERCENTAGES-WILL BE FILLED IN AFTERWARDS FROM PREVIOUS QUESTIONS)

	1988
HMO	_____ %
PPO	_____ %
Traditional fee-for-service, without pre-admission certification	_____ %
Fee-for-service with pre-admission certification	_____ %
Total	100%

Some health plans are using payment system changes to control the number and cost of physicians' services. Has your company implemented the following changes?

14. Changing the coding of physician services?

YES.....1 (Go to Q14A)
NO.....2 (Go to Q15)
DK.....9 (Go to Q15)

A) Have you formally evaluated its effectiveness?

NA.....0
YES.....1 (Go to Q14B)
NO.....2 (Go to Q14D)
DK.....9 (Go to Q15)

B) What has been the effect on total cost?

NA.....0
INC.....1 (Go to Q14Bi)
DEC.....2 (Go to Q14Bii)
NOCHANGE.....3 (Go to Q14C)
DK.....9 (Go to Q14C)

i) What is the percent increase?_____%

ii) What is the percent decrease?_____%

C) What has been the effect on utilization?

NA.....0
INC.....1 (Go to Q14Ci)
DEC.....2 (Go to Q14Cii)
NOCHANGE.....3 (Go to Q15)
DK.....9 (Go to Q15)

i) What is the percent increase?_____% (Go to Q15)

ii) What is the percent decrease?_____% (Go to Q15)

D) i) What is your overall feeling toward the effectiveness of changing the coding of physician services?

ii) How do you think changing the coding of physician services has effected cost and utilization outside the hospital?

15. (And have you implemented) "bundling" of physician services into broader categories than visits or procedures, such as a single payment per episode or illness

YES.....1 (Go to Q15A)
NO.....2 (Go to Q16)
DK.....9 (Go to Q16)

A) Have you formally evaluated its effectiveness?

NA.....0
YES.....1 (Go to Q15B)
NO.....2 (Go to Q15D)
DK.....9 (Go to Q16)

B) What has been the effect on total cost?

NA.....0
INC.....1 (Go to Q15Bi)
DEC.....2 (Go to Q15Bii)
NOCHANGE.....3 (Go to Q15C)
DK.....9 (Go to Q15C)

i) What is the percent increase?_____%

ii) What is the percent decrease?_____%

C) What has been the effect on utilization?

NA.....0
INC.....1 (Go to Q15Ci)
DEC.....2 (Go to Q15Cii)
NOCHANGE.....3 (Go to Q16)
DK.....9 (Go to Q16)

i) What is the percent increase?_____% (Go to Q16)

ii) What is the percent decrease?_____% (Go to Q16)

D) i) What is your overall feeling toward the effectiveness of "bundling" of physician services?

ii) How do you think "bundling" of physician services has effected cost and utilization outside the hospital?

16. (And have you implemented) preventing multiple physicians from billing for the same service, such as eliminating payments for assistant surgeon?

YES.....1 (Go to Q16A)
NO.....2 (Go to Q17)
DK.....9 (Go to Q17)

A) Have you formally evaluated its effectiveness?

NA.....0
YES.....1 (Go to Q16B)
NO.....2 (Go to Q16D)
DK.....9 (Go to Q17)

B) What has been the effect on total cost?

NA.....0
INC.....1 (Go to Q16Bi)
DEC.....2 (Go to Q16Bii)
NOCHANGE.....3 (Go to Q16C)
DK.....9 (Go to Q16C)

i) What is the percent increase?_____%

ii) What is the percent decrease?_____%

C) What has been the effect on utilization?

NA.....0
INC.....1 (Go to Q16Ci)
DEC.....2 (Go to Q16Cii)
NOCHANGE.....3 (Go to Q17)
DK.....9 (Go to Q17)

i) What is the percent increase?_____% (Go to Q17)

ii) What is the percent decrease?_____% (Go to Q17)

D) i) What is your overall feeling toward the effectiveness of preventing multiple billing?

ii) How do you think preventing multiple billing has effected cost and utilization outside the hospital?

17. (IF COMPANY HAS IMPLEMENTED CHANGES IN PAYMENT SYSTEM, ASK:) How do you evaluate the payment system changes to determine if they are effective?

18. (IF COMPANY DOES NOT CURRENTLY USE ONE OR MORE OF THE PAYMENT SYSTEM CHANGES MENTIONED ABOVE, ASK:) Please indicate any reasons you may have for not implementing payment system changes. (PROBE)

19. Do you own any third party administrators?

YES.....1
NO.....2
DK.....9

20. Do you own any utilization review organizations?

YES.....1
NO.....2
DK.....9

21. Have you entered into any joint ventures or have an equity position with the following organizations during 1987-88, outside your PPO or HMO business? (Joint ventures are defined as contractual arrangements where the two parties share in the profit and/or losses) (READ CATEGORIES AND CIRCLE APPROPRIATE RESPONSE)

	<u>YES</u>	<u>NO</u>	<u>DK</u>
a) Third party administrators.....	1	2	9
b) Utilization review organizations.....	1	2	9
c) Other insurers.....	1	2	9
d) Case management companies.....	1	2	9
e) Data analysis firms.....	1	2	9
f) Software firms.....	1	2	9
g) Consulting firms.....	1	2	9
h) continuing care retirement communities.....	1	2	9
i) other			
(specify) _____			

22. Have you purchased any of the following organizations during 1987-88? (READ CATEGORIES AND CIRCLE APPROPRIATE RESPONSE)

	<u>YES</u>	<u>NO</u>	<u>DK</u>
a) Third party administrators.....	1	2	9
b) Utilization review organizations.....	1	2	9
c) Other insurers.....	1	2	9
d) Hospital chains.....	1	2	9
e) Physician groups.....	1	2	9
f) Case management companies.....	1	2	9
g) Data analysis firms.....	1	2	9
h) Software firms.....	1	2	9
i) Consulting firms.....	1	2	9
j) continuing care retirement communities.....	1	2	9
k) other			
(specify) _____			

(REFER TO INTRODUCTION QUESTION 1 AND GO TO NEXT APPROPRIATE SECTION)

CMS LIBRARY



3 8095 00014076 0